Fraunhofer
IAF

FRAUNHOFER-INSTITUT FÜR ANGEWANDTE FESTKÖRPERPHYSIK IAF

KOMPETENZZENTRUM QUANTENCOMPUTING BADEN-WÜRTTEMBERG

VERBUNDFORSCHUNGSPROJEKT

QUANTENOPTIMIERUNG MIT RESILIENTEN ALGORITHMEN (QORA)

Abschlussbericht
März 2023

Koordinator:
Dr. Thomas Wellens
Fraunhofer-Institut für Angewandte Festkörperphysik IAF, Freiburg

Projektnummer: 3-4332-62-IAF/9

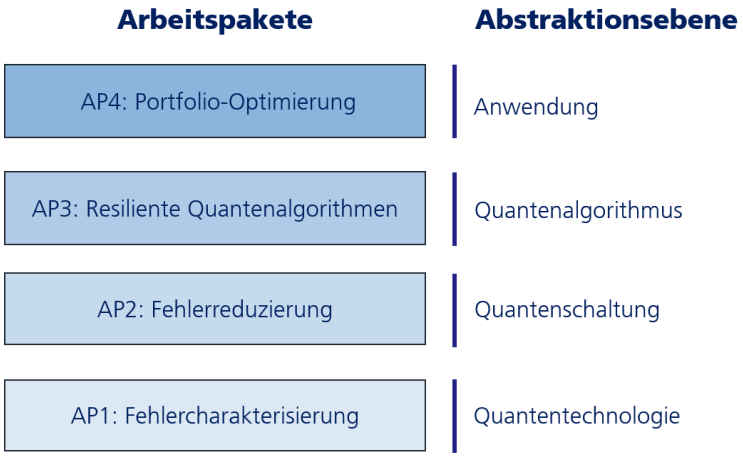
Projektpartner:

- Prof. Gerhard Hellstern, Duale Hochschule Baden-Württemberg, Ravensburg (DHBWR)
- Prof. Guido Burkard, Fachbereich Physik, Universität Konstanz (UKON)
- Prof. Daniel Braun, Institut für Theoretische Physik, Eberhard-Karls-Universität Tübingen (EKUT)
- Prof. Ilia Polian, Institut für Technische Informatik, Universität Stuttgart (USTUTT-ITI) und Prof. Stefanie Barz, Institut für Funktionelle Materie und Quantentechnologien, Universität Stuttgart (USTUTT-FMQ)

Inhalt

1 Zusammenfassung

Ziel von QORA war die Entwicklung von fehlerresistenten Quantenalgorithmen für Optimierungsprobleme, insbesondere im Bereich Portfoliooptimierung, die auf Quantenhardware mit nichttrivialen Fehlerraten lauffähig sind. Hierfür wurde ein schichtenübergreifender Ansatz auf mehreren Ebenen verfolgt, welche den vier Arbeitspaketen entsprechen:



Grundlage des Arbeitsprogramms war zunächst die Fehlercharakterisierung (AP1) auf dem IBM-Quantencomputer in Ehningen (im Folgenden als „ibmq_ehningen“ bezeichnet). Darauf aufbauend wurden Methoden zur Fehlerreduzierung (AP2) sowie resiliente Quantenalgorithmen entwickelt (AP3). Diese Methoden waren die Grundvoraussetzung für die Nutzung von ibmq_ehningen zur Lösung von Optimierungsproblemen im Bereich von Portfolios (AP4). Mit den heutigen Quantencomputern (inkl. ibmq_ehningen) können dabei nur relativ kleine und daher noch nicht anwendungsrelevante Probleminstanzen behandelt werden. In Anbetracht der in den nächsten Jahren zu erwartenden Entwicklungen wollten wir mit QORA zur Schaffung der nötigen wissenschaftlichen Grundlagen beitragen, um eine möglichst frühzeitige anwendungsrelevante Nutzung von Quantencomputern zu ermöglichen.

Um einen Kurzeinblick in den Verlauf des Projekts zu geben, stellen wir diesen zunächst anhand der Meilensteine dar.

M1	einfaches Fehlermodell (ohne Korrelationen) bereit	AP1	Monat 8	erfüllt
M2	genaueres Fehlermodell auf Basis der Ergebnisse von AP1.3-4 bereit	AP1	Monat 16	erfüllt
M3	funktionaler Test entwickelt und implementiert	AP1	Monat 24	erfüllt
M4	lauffähige Resilienz-Bewertungs- und Steigerungsverfahren unter Annahme einfacher Fehler	AP2	Monat 12	weitgehend erfüllt
M5	auf das genauere Fehlermodell aus AP1 zugeschnittene verbesserte Simulations-, Fehlerkorrektur-/erkennungs- und Transpilationsverfahren	AP2	Monat 24	erfüllt
M6	Weiterentwicklung und Optimierung von QAOA beendet	AP3	Monat 8	erfüllt

M7	gradientenfreie Optimierungsverfahren auf QAOA übertragen und getestet	AP3	Monat 12	erfüllt
M8	fehlerresistenter QAOA-Algorithmus für Anwendung bereit	AP3	Monat 24	erfüllt
M9	Einsatz der Prototypen inkl. der Implementierung auf ibmq_ehningen	AP4	Monat 12	erfüllt
M10	Demonstration der Verbesserungen für die Portfoliooptimierung unter Berücksichtigung der Ergebnisse AP1-AP3, Abschätzung eines Quantenvorteils	AP4	Monat 24	erfüllt

Der erste Meilenstein (M1) in **AP1** (Fehlercharakterisierung) zielte auf die Erstellung eines detaillierten Modells der auf ibmq_ehningen auftretenden Fehler ab, wobei zuerst jedes Quantengatter einzeln betrachtet wurde. Da sich das in der Literatur beschriebene Verfahren der Gattersatztomographie („gate set tomography") hier für die 2-Qubit-Gatter als nicht praktikabel erwies, mussten wir eine neue Methode entwickeln (siehe Kap. 2.1.1). Auch damit ist jedoch die ursprünglich angestrebte detaillierte Charakterisierung mit hohem Aufwand verbunden, weswegen wir für die anderen Arbeitspakete zunächst auf das von IBM zur Verfügung gestellte weniger realistische Fehlermodell zurückgegriffen haben. Eine weiterführende Charakterisierung der Korrelationen zwischen gleichzeitig durchgeführten Gattern, welche auf Crosstalk-Effekte zurückzuführen sind, wurde dann unter Ausnutzung paralleler „Randomized Benchmarking" Experimente durchgeführt (siehe Kap. 2.1.2). Die erhaltenen Daten aus dieser Charakterisierung erlaubten uns schließlich, ein genaueres Crosstalk-empfindliches Fehlermodell zu definieren und damit den zweiten Meilenstein (M2) zu erreichen (siehe Kap. 2.1.3). Begleitend zu diesen Arbeiten wurden auf ibmq_ehningen stark verschränkte Zustände von mehreren Qubits erzeugt und durch geeignete Fidelity-Messungen sowie Bell-Experimente charakterisiert (siehe Kap. 2.1.4). Des Weiteren wurde der Einfluss zeitlich korrelierter Fehler im QAOA-Algorithmus erforscht. Hier wurde ein bislang noch unbekannter Mechanismus entdeckt, mit Hilfe dessen der Quantenalgorithmus diese Fehler teilweise kompensieren kann (siehe Kap. 2.1.5). Außerdem wurden funktionale Test (M3) für die effiziente Charakterisierung von Gattern entwickelt und für 1-Qubit Gatter implementiert (siehe Kap. 2.2.6).

In **AP2** (Fehlerreduzierung) haben wir uns von den ursprünglich vorgesehenen drei Ansätzen (Quantenfehlerkorrektur, Fehlererkennung und Transpilation) auf einen (Transpilation) konzentriert. Gründe hierfür waren einerseits Schwierigkeiten bei der Anwerbung wissenschaftlichen Fachpersonals. Andererseits kamen wir auch zu dem Schluss, dass die gegenwärtige Hardware (ibmq_ehningen) für die Durchführung von Quantenfehlerkorrektur noch nicht geeignet ist. Der Aspekt der Fehlerreduzierung durch Transpilation wurde hingegen erfolgreich bearbeitet. Neben der sehr guten Transpilationsergebnisse (s. 2.2 für eine Diskussion im Detail) wurden neuartige Verfahren vorgeschlagen, die Kalibrationsdaten, Struktureigenschaften des Algorithmus und Besonderheiten sog. bipotenter Architekturen gewinnbringend einsetzen. Somit kann Meilenstein M4 als weitgehend erfüllt gelten. Die Integration der Fehlermodelle aus AP1 in die Qiskit Simulationssoftware (Kap. 2.2.2) und die Transpilation unter der Annahme von *Crosstalk*-Fehlern (Kap. 2.2.3) wurde durchgeführt. Damit wurde ebenfalls Meilensteins M5 erfüllt.

In **AP3** (resiliente Quantenalgorithmen) haben wir verschiedene Varianten des QAOA-Algorithmus zunächst auf dem fehlerfreien Simulator implementiert und miteinander verglichen. Zwei bestimmte Varianten (davon eine von uns selbst Entwickelte) haben sich für die Lösung des Portfoliooptimierungsproblems im fehlerfreien Fall als besonders geeignet erwiesen (Meilenstein M6, siehe Kap. 2.3.1). Andererseits stellten sich gerade diese Varianten als besonders empfindlich gegenüber Fehlern heraus (siehe Kap. 2.3.4). Parallel dazu wurden verschiedene klassische Verfahren zum Auffinden der optimalen Parameter des QAOA-Schaltkreises untersucht und diejenigen identifiziert, die unter

verschiedenen Voraussetzungen (mit oder ohne Berücksichtigung der durch eine endliche Zahl von Messungen bedingten statistischen Schwankungen) die besten Ergebnisse liefern (Meilenstein M7, siehe Kap. 2.3.2). Außerdem wurden leichte und schwere Portfolioprobleme durch eine Analyse der Verteilung der Korrelationen sowie Gewinne der Aktien identifiziert. Als letztes haben wir eine Reihe von Fehlerreduzierungs- und Fehlermitigationsmethoden, basierend auf den Ergebnissen aus AP1 und AP2, implementiert und damit einen fehlerresistenteren QAOA-Algorithmus formuliert (Meilenstein M8) und implementiert (Kap. 2.4.2).

In **AP4** (Portfoliooptimierung) haben wir zunächst anhand realer Markdaten von DAX-Aktien verschiedene Instanzen von Portfoliooptimierungsproblemen definiert, welche wir in den anderen Arbeitspaketen als Benchmark für unsere Algorithmen benutzen konnten. Diese prototypischen Instanzen haben eine klare mathematische Struktur (als sogenanntes „QUBO"-Problem), enthalten aber einige vereinfachende Annahmen (z.B. Beschränkung auf binäre Portfoliogewichte und Vernachlässigung von Transaktionskosten). Aus diesem Grund haben wir auch untersucht, wie diese Annahmen realitätsnäher gestaltet werden können, ohne die grundlegende Struktur des Problems zu ändern (siehe Kap. 2.4.1). Die Lösung dieser prototypischen Probleme auf Simulatoren sowie ibmq_ehningen wurde umgesetzt (Meilenstein M9, siehe Kap. 2.4.2). Innerhalb dieses Arbeitspakets fand auch ein Austausch mit den assoziierten Partnern statt. Obwohl die Partner grundsätzlich Interesse an unserem Projekt bekundeten, stellte sich dabei heraus, dass speziell das Portfoliooptimierungsproblem für die Bausparkassen nicht von herausragender Dringlichkeit ist. Aus diesem Grund haben wir begonnen, auch andere Use Cases (wie z.B. die Merkmalsauswahl bei Klassifikationsalgorithmen für Kreditscoring) zu identifizieren, auf welche sich aufgrund ihrer ähnlichen Struktur die in QORA entwickelten Techniken anwenden lassen (siehe Kap. 2.4.1). In der zweiten Projekthälfte wurden außerdem weitere Fehlerreduzierungsmethoden, basierend auf den Resultaten aus AP1-3, implementiert, welche die bereits erhaltenen Ergebnisse (des Simulators und von ibmq_ehningen) weiter verbessert haben. Abschließend wurde auch eine Abschätzung des Quantenvorteils vorgenommen (Meilenstein M10), bei der die Performanz des QAOA-Algorithmus mit leistungsstarken klassischen Algorithmen verglichen wurde. Zusammenfassend, konnten wir auf Basis dieser Analyse keine Hinweise auf einen Quantenvorteil finden.

Insgesamt war das Projekt QORA sehr hilfreich, um nötige Kompetenzen im Bereich Quantencomputing in Baden-Württemberg aufzubauen. Das zum Schluss genannte Ergebnis, dass für den Fall der Portfoliooptimierung bislang kein Quantenvorteil gefunden wurde, mag zwar auf den ersten Blick ernüchternd scheinen. Die Identifizierung von für Quantencomputing geeigneten Anwendungsfällen mit überzeugendem Nachweis eines Quantenvorteils ist jedoch eine schwierige Herausforderung für aktuelle Forschung, der wir uns auch in der zweiten Förderperiode (QORA II) stellen werden, um Unternehmen von den Potentialen des Quantencomputings zu überzeugen und diese für sie nutzbar zu machen.

2 Überblick der wissenschaftlichen Arbeiten

Im Folgenden geben wir einen Überblick der wissenschaftlichen Arbeiten, wobei wir die wichtigsten Projektergebnisse sowie gegebenenfalls Abweichungen vom geplanten Projektverlauf diskutieren.

2.1 AP 1: Fehlercharakterisierung

Grundlage des QORA-Arbeitsprogramms bildet die Fehlercharakterisierung. Die Leitidee dabei ist, dass eine genauere Kenntnis der Fehler, die auf dem Quantencomputer auftreten, hilfreich ist, um diese besser bekämpfen zu können. Da jeder Quantenschaltkreis sich in 1-Qubit- und 2-Qubit-Gatter zerlegen lässt, hat IAF zunächst die Eigenschaften dieser elementaren Gatter charakterisiert (siehe Kap. 2.1.1 und 2.1.2), wobei wir nachgewiesen haben, dass im Fall der IBM Quantum-Systeme insbesondere Crosstalk-Effekte eine wichtige Rolle spielen (Kap. 2.1.2). Diese Arbeiten mündeten in die Erstellung eines verbesserten Fehlermodells, welches über das von IBM in Qiskit bereitgestellte Modell hinausgeht (Kap. 2.1.3). Parallel dazu hat USTUTT-FMQ die Fehlereigenschaften des Systems anhand der Erzeugung verschränkter Zustände charakterisiert (Kap. 2.1.4). Diese der Charakterisierung der gegebenen IBM-Hardware dienenden Arbeitspakete wurden von zwei eher grundlagenorientierten Arbeiten begleitet, von denen eine die Möglichkeit der Adaptivität des QAOA-Algorithmus gegenüber zeitlich korrelierten Fehlern aufzeigt (UKON, Kap. 2.1.5), während die andere sich mit Prinzipien des funktionalen Tests von Quantenchips beschäftigt (EKUT, Kap. 2.1.6).

2.1.1 Gattercharakterisierung ohne Korrelationen (AP1.1)

In Arbeitspaket AP1.1 wurde eine Charakterisierung aller 1- und 2-Qubit-Gatteroperationen auf verschiedenen IBMQ-Systemen mit 27 Qubits (u.a. ibmq_ehningen) durchgeführt. Dabei haben wir uns zunächst auf die Annahme beschränkt, dass Gatterfehler ohne Korrelationen zwischen verschiedenen Qubits auftreten (siehe auch „Korrelationen zwischen gleichzeitig durchgeführten Gattern und Crosstalk-Effekte“ weiter unten). Im Fall der 1-Qubit-Gatteroperationen konnten wir hier auf das bekannte Verfahren der Gattersatztomographie („gate set tomography“ [Nielsen2021: E. Nielsen *et. al.*, Quantum **5**, 557 (2021)]) zurückgreifen. Letzteres erlaubt eine präzise Bestimmung aller 12 Fehlerparameter (darunter die mittlere Fehlerrate) eines jeden Gatters einer vorgegebenen Menge von 1-Qubit-Gattern (siehe Abb. 1.1). Eine geeignete Modifizierung der Gattertomographie durch Fixierung eines als fehlerfrei angenommenen Anfangszustands erlaubte uns außerdem, eine Charakterisierung der Auslesefehler der einzelnen Qubits durchzuführen (siehe Abb. 1.2). Alle auf diese Weise ermittelten Fehlerraten stimmen bei Berücksichtigung der vorhandenen statistischen Fluktuationen gut mit den von IBM bereitgestellten Kalibrierungsdaten überein.

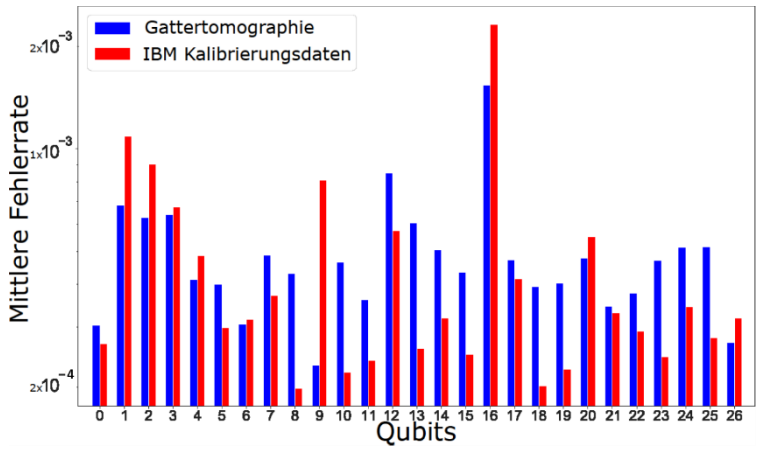


Abbildung 1.1: Mittlere Fehlerrate der mittels Gattertomographie mit 784 Charakterisierungsschaltkreisen und 8192 Shots pro Schaltkreis rekonstruierten 1-Qubit Gatteroperationen für alle 27 Qubits von ibmq_ehningen im Vergleich zu den IBM-Kalibrierungsdaten vom 15. November 2021.

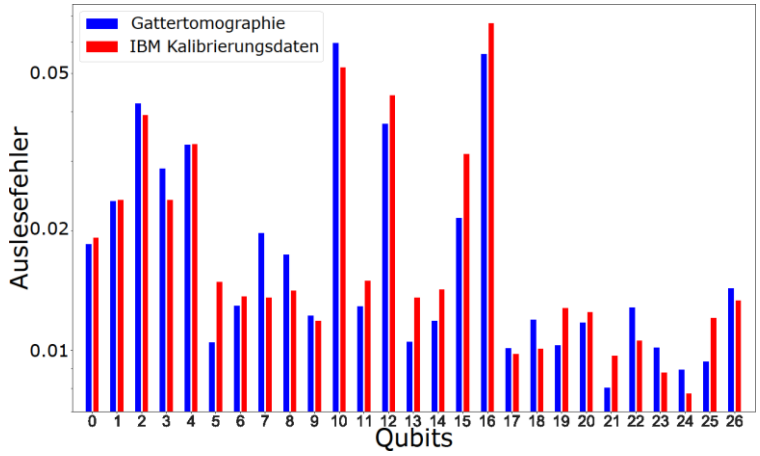


Abbildung 1.2: Auslesefehler der wie in Abbildung 1 rekonstruierten 1-Qubit-Messoperatoren für alle 27 Qubits von ibmq_ehningen im Vergleich zu den IBM-Kalibrierungsdaten vom 15. November 2021.

Im zweiten Schritt haben wir eine Charakterisierung der deutlich fehleranfälligeren 2-Qubit-CNOT-Operationen durchgeführt. Hierfür wurde jedoch nicht auf das Verfahren der Gattersatztomographie zurückgegriffen, weil diese für eine vollständige Charakterisierung aller 240 2-Qubit-Fehlerparameter auf allen möglichen Qubit-Paaren eines 27-Qubit-Chips zu zeitaufwendig ist. Stattdessen haben wir im Zuge des Projekts ein neues Verfahren für die Prozess-Tomographie der CNOT-Gatteroperationen entwickelt. Dieses Verfahren bezieht die im ersten Schritt rekonstruierten 1-Qubit Operationen mit ein, indem effektive Zustandspräparationen und -messungen durch geeignete „fiducial" Schaltkreise, bestehend aus 1-Qubit Operationen, realisiert werden. Ähnlich wie die Gattersatztomographie beruht unser Verfahren auch auf der wiederholten Anwendung des zu charakterisierenden Gatters, welches eine Verstärkung des verursachten Gesamtfehlers und damit eine präzisere Bestimmung der Gatterfehler erlaubt. Die Wiederholung der jeweiligen Gatter erfolgt hierbei innerhalb sog. Germ-Schaltkreise, zusammen mit weiteren 1-Qubit Operationen, um eine möglichst effiziente Verstärkung aller Fehlerparameter zu erreichen. Für die Erstellung der Germ-Schaltkreise wurde ein geeignetes, in [Nielsen2021] beschriebenes, Optimierungsverfahren implementiert.

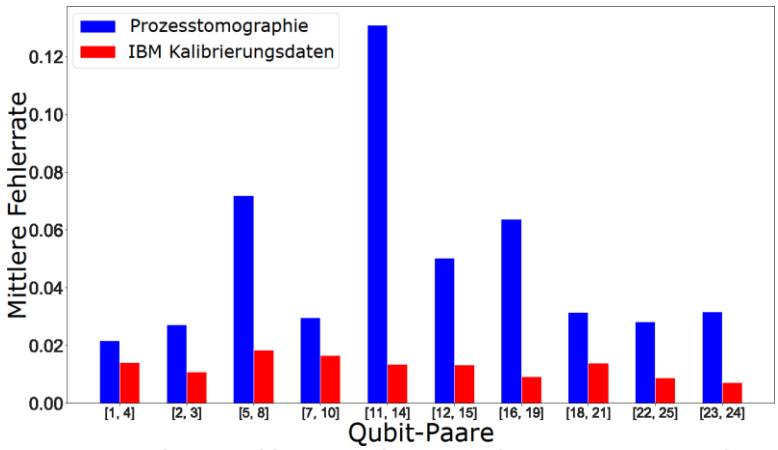


Abbildung 1.3: Mittlere Fehlerrate der mittels Prozesstomographie mit 3024 Charakterisierungsschaltkreisen und 8192 Shots pro Schaltkreis rekonstruierten CNOT-Gatter bei gleichzeitiger Charakterisierung einer beispielhaften Menge an Qubit-Paaren im Vergleich zu den IBM Kalibrierungsdaten vom 24. November 2021.

Die Rekonstruktion der fehlerhaften CNOT-Gatter erfolgt schlussendlich über das Anpassen einer vollständig positiv parametrisierten Gatteroperation an die aus den Germ-Schaltkreisen erhaltenen Daten mit Hilfe einer Optimierung der zugrundeliegenden χ^2-Statistik. Wir haben diese Charakterisierung für alle Paare von Qubits durchgeführt, wobei aufgrund der Topologie des Quantenchips höchstens zehn Paare gleichzeitig charakterisiert werden können (siehe Abb. 1.3). Mit den erhaltenen Daten für die 1- und 2-Qubit-Gatter ist es möglich, ein einfaches Fehlermodell (ohne Korrelationen) zu erstellen (Meilenstein M1). Es zeigt sich jedoch, dass die Annahme der Korrelationsfreiheit zwischen verschiedenen Qubits im Allgemeinen nicht erfüllt und eine genauere Charakterisierung dieser Crosstalk-Effekte notwendig ist.

2.1.2 Korrelationen zwischen gleichzeitig durchgeführten Gattern und Crosstalk-Effekte (AP1.3)

Hierfür haben wir systematisch die mittleren CNOT-Fehlerraten aller Qubit-Paare mittels *Randomized Benchmarking* [1] für folgende Fälle untersucht: (1) Auf den Nachbar-Qubits werden keine parallelen Operationen durchgeführt, (2) auf einem Nachbar, oder auf (3) mehreren Nachbarn werden gleichzeitig Charakterisierungen durchgeführt. *Randomized Benchmarking* ist in diesem Fall das effizientere Protokoll, da wir uns in erster Linie für die mittleren Fehlerraten interessieren und eine Rekonstruktion aller Fehlerparameter dafür nicht notwendig ist. Abbildung 1.4 zeigt das Ergebnis einer Crosstalk-Untersuchung für ein exemplarisch ausgewähltes Qubit-Paar. Zusammenfassend findet man, dass der zu beobachtende Crosstalk-Effekt meistens auf ein spezielles benachbartes Qubit-Paar zurückzuführen ist und nicht durch die gleichzeitige Ausführung von mehr als zwei CNOT-Gattern verursacht wird (siehe Abb. 1.4). Für die Charakterisierung der kompletten Korrelationsmatrix, also der Fehlerraten aller Qubit-Paare bei gleichzeitiger Charakterisierung der Nachbar-Qubits, ist es also ausreichend, sich auf jeweils eines der nächsten Nachbarpaare zu beschränken. Wir haben diese Korrelationsmatrix für ibmq_ehningen sowie ibmq_toronto erstellt und den Projektpartnern zur Verfügung gestellt.

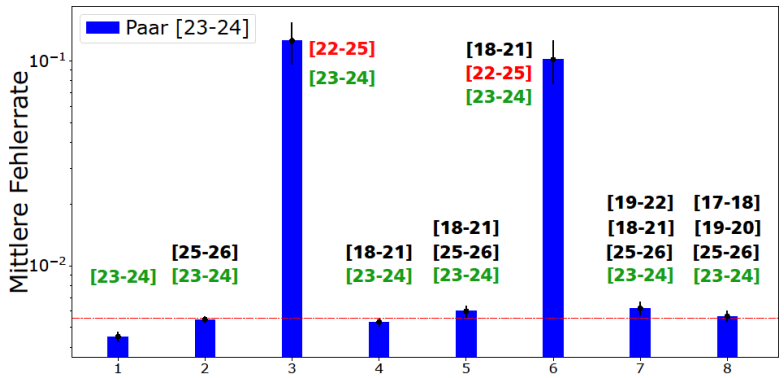


Abbildung 1.4: Mittlere Fehlerrate des auf dem Qubit-Paar [23-24] ausgeführten CNOT-Gatters bei gleichzeitiger Charakterisierung der angezeigten benachbarten Qubit-Paare. Gleichzeitiges Ausführen des CNOT-Gatters [22-25] führt zu einem signifikanten Anstieg der Fehlerrate des CNOT-Gatters [23-24]. Die rote gestrichelte Linie stellt den von IBM angegebenen Fehler des CNOT-Gatters [23-24] dar.

Bei einer weitergehenden Untersuchung der in Abb. 1.4 beobachteten Crosstalk-Effekte haben wir herausgefunden, dass die ansteigenden Fehlerraten nicht nur bei der parallelen Ausführung von 2-Qubit Gattern auftreten. Stattdessen können 2-Qubit Gatter generell Auswirkungen auf benachbarte Qubits haben, wenn eine sog. Frequenz-Kollision vorliegt. Aufgrund einer Resonanz werden hierbei ungewollte Übergänge innerhalb eines Qubits oder zweier benachbarter Qubits angeregt. In Abb. 1.5 haben wir zwei solche Frequenz-Kollisionen und die daraus folgenden Resonanzen auf dem IBMQ Ehningen System untersucht und deren Auswirkungen mit zwei gezielt entworfenen Schaltkreisen gemessen. In beiden Fällen führen ungewollt getriebene Resonanzen zu einer Population des energetisch höheren 2-Zustands von einem der Transmon-Qubits, was wiederum ein stark fehlerbehaftetes Messergebnis zur Folge hat.

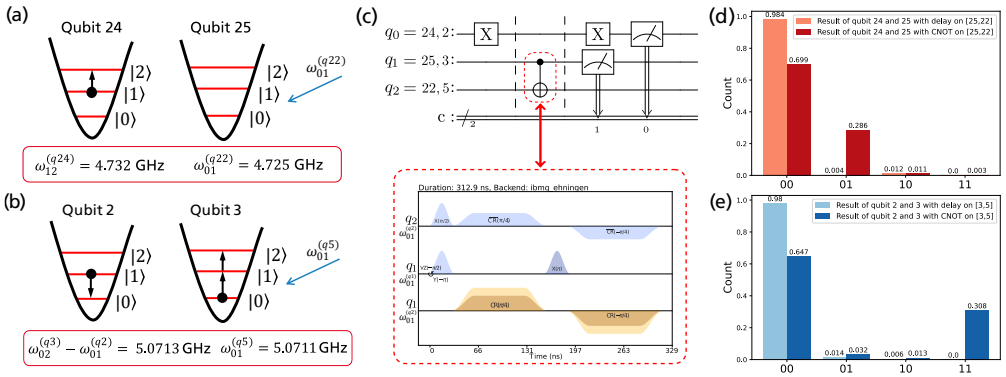


Abbildung 1.5: (a,b) Illustration von zwei Frequenz-Kollisionen auf IBMQ Ehningen. Die Transmon-Qubits sind mit den 3 niedrigsten Energiezuständen schematisch dargestellt, während die blauen Pfeile CR (Cross-Resonance)-Pulse repräsentieren, welche bei der Anwendung von CNOT-Gattern auftreten. (c) Schaltkreis, um die Resonanzen (a) und (b) gezielt zu treiben. Die rote Box zeigt das Mikrowellenpulsschema des CNOT-Gatters. (d,e) Histogramme der Messergebnisse des auf den in (a) und (b) dargestellten Qubits ausgeführten Schaltkreises (c). Zum Vergleich zeigen die Histogramme auch die Resultate des Schaltkreises (c), bei dem anstatt eines CNOT-Gatters eine Wartezeit derselben Länge eingefügt wird. Der durch das CNOT-Gatter hervorgerufene Crosstalk-Effekt äußert sich in einer Abweichung vom erwarteten fehlerfreien Messergebnis „00".

Um die Auswirkungen von Resonanzprozessen, wie in Abb. 1.5 gezeigt, auf IBMQ Ehningen systematischer auszumessen, haben wir ein neues Charakterisierungsprotokoll entworfen. Dieses besteht aus simultanen *Randomized-Benchmarking*-Experimenten auf ausgewählten Qubit-Dreiergruppen, bei dem immer ein Qubit-Paar parallel mit einem zusätzlichen benachbarten Qubit charakterisiert wird. Die Qubit-Dreiergruppen werden

dabei nach folgendem Schema definiert: Wir wählen zuerst ein Qubit-Paar aus und markieren anschließend das Qubit, auf das der CR-Puls angewandt wird, welcher dem auf diesem Paar ausgeführten CNOT-Gatter zugrundeliegt. Anschließend fügen wir aus den möglichen Nachbarn dieses Qubits eines zu dem ausgewählten Paar hinzu. Dieses Schema führt zu insgesamt 41 Dreiergruppen von Qubits, wobei die gemessenen 1- und 2-Qubit-Fehlerraten jeweils die Crosstalk-Effekte innerhalb dieser Gruppen wiederspiegeln.

Abbildung 1.6 zeigt die am 23. Januar 2022 gemessenen Fehlerraten und vergleicht diese mit den entsprechenden IBM-Kalibrierungsdaten zum selben Zeitpunkt. Die Charakterisierungsdaten zeigen, dass sich Crosstalk-Effekte auf bestimmte Qubitgruppen beschränken, wie zum Beispiel auf die in Abb. 4 diskutierten Qubits, während viele der Qubit-Dreiergruppen keine aufällige Erhöhung der Fehlerraten zeigen. Die Erhöhung der Fehlerraten im ersten Fall ist allerdings signifikant und kann zu extrem reduzierten Güten bei der Ausführung von bestimmten Schaltkreisen führen. Es ist daher von Interesse, die neu gewonnenen Informationen über besagte Crosstalk-Effekte bei der Auswahl der Hardware-Qubits und bei der Transpilation der untersuchten Schaltkreise zu berücksichtigen. Dies kann mit Hilfe eines geeigneten Crosstalk-sensitiven Fehlermodels realisiert werden.

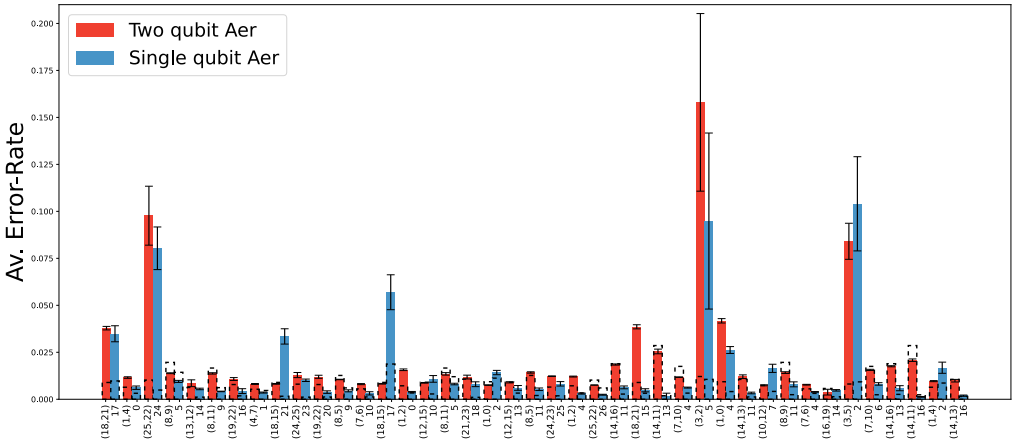


Abbildung 1.6: Parallel charakterisierte mittlere Fehlerraten von 1-Qubit- (blau) und 2-Qubit-Gattern (rot), wobei alle 41 Qubit-Dreiergruppen am 23. Januar 2022 auf ibmq_ehningen vermessen wurden. Die Charakterisierung wurde mittels Randomized Benchmarking mit den Schaltkreislängen [1, 3, 10, 20, 40, 65, 95, 130, 175] und einer fünfmaligen Wiederholung jedes Schaltkreises mit 10 000 Shots pro Schaltkreis durchgeführt. Alle Schaltkreise enthalten geeignet platzierte Barrieren, um ein paralleles Ausführen der 1- und 2-Qubit Gatter sicherzustellen. Gestrichelte Boxen entsprechen den zugehörigen IBM-Kalibrierungsdaten am selben Tag.

2.1.3 Erstellung eines Fehlermodels mit Crosstalk-Effekten (AP1.3, AP1.5)

Standardmäßige Fehlermodelle, wie sie zum Beispiel von IBM für jedes IBMQ-System zur Verfügung gestellt werden, beziehen Crosstalk-Effekte nicht mit ein, sondern versehen jedes Gatter eines Schaltkreises einfach mit einem thermischen Relaxationsfehler und einem Depolarisationsfehler. Letztere werden anhand der Kalibrierungsdaten des jeweiligen Systems ausgewählt, also der mittleren Fehlerraten der Gatter und der thermischen Relaxations- und Dekohärenzzeiten der Qubits. Im Zuge des Projektes QORA haben wir ein Fehlermodell entwickelt (Meilenstein M2), welches über diese Art von Fehlermodell hinausgeht und Crosstalk-Effekte berücksichtigt. Hierzu verfolgen wir die folgende einfache Strategie. Zunächst überfliegen wir alle 2-Qubit-Gatter eines gegebenen transpilierten Schaltkreises und überprüfen, ob deren Nachbarqubits „aktiv“ sind, also ob schon ein Gatter auf ihnen ausgeführt wurde. Wenn ein Nachbarqubit aktiv ist, versehen wir das entsprechende 2-Qubit-Gatter mit einem Depolarisationsfehler

entsprechend der in Abb. 1.6 (rote Balken) dargestellten Fehlerrate. Wenn es mehr als ein aktives Nachbarqubit gibt, nehmen wir den entsprechend größeren Fehler, oder den Fehler aus den Kalibrierungsdaten, falls es keinen aktiven Nachbarn gibt. Zusätzlich wenden wir auf die aktiven Nachbarqubits eine Identitätsoperation mit der in Abb. 1.6 (blaue Balken) gezeigten Fehlerrate an, um den Crosstalk-Effekt zu simulieren (siehe Abb.1.7). Zum Schluss fügen wir dem Schaltkreis noch zusätzliche Amplituden- und Phasendämpfungs-Kanäle hinzu, um durch Wartezeiten zwischen den Gatterpulsen verursachten Dekohärenz Effekten Rechnung zu tragen. Durch die Bereitstellung dieses erweiterten, genaueren Fehlermodells erreichten wir das Ziel des Meilensteins M2. Die weitere Integration dieses Fehlermodells in den Simulator und deren Validierung durch Experimente auf IBMQC sind Resultate des Arbeitspakets AP2.4 und wurden außerdem in dem Preprint [8] veröffentlicht.

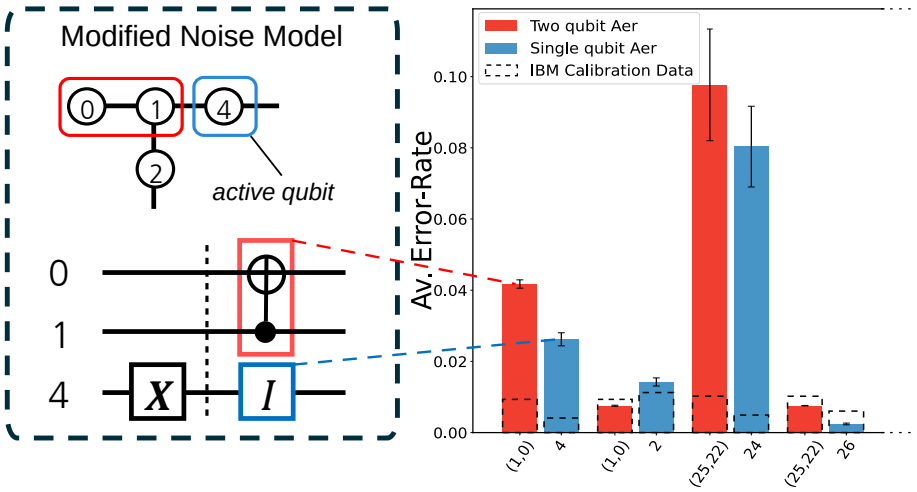


Abbildung 1.7: Schematische Darstellung des Crosstalk-empfindlichen Fehlermodells. Immer wenn ein CNOT Gatter auf ein Qubit-Paar angewendet wird (z.B. [0,1]), welches in der direkten Nachbarschaft eines aktiven Qubits liegt, versehen wir es mit einem Depolarisationsfehler mit der entsprechenden Stärke, welche aus den Charakterisierungsdaten in Fig. 1.6 hervorgeht. Gleichzeitig fügen wir dem aktiven Qubit ein zusätzliches Identitäts-Gatter mit entsprechendem Depolarisationsfehler hinzu, um den durch Crosstalk verursachten Korrelationen Rechnung zu tragen.

Zusammenfassend stellen wir fest, dass Crosstalk-Effekte einen wesentlichen Anteil an den tatsächlich beobachteten Fehlern von supraleitenden Transmon-Prozessoren darstellen, welcher sich nicht in den Kalibrierungsdaten des Herstellers wiederspiegelt. Frequenz-Kollisionen sind dabei eine Hauptursache für solche Crosstalk-Effekte, die sich in Korrelationen zwischen gleichzeitig ausgeführten Gattern auswirken. Die im Projekt QORA entwickelten Methoden für die Charakterisierung und Simulation von Crosstalk-Effekten leisten also einen wichtigen Beitrag zum besseren Verständnis der Arbeitsweise der untersuchten Prozessoren und bieten einen Startpunkt für die Entwicklung besserer fehlerresistenter Algorithmen. Letztere können zum Beispiel aus besseren Transpilations- oder Fehlermitigationsverfahren, welche in AP.2.2 untersucht wurden, hervorgehen. Des Weiteren bestärken uns die in AP1 gefundenen Ergebnisse darin, dass eine physikalisch motivierte Charakterisierung des Quantenprozessors, wie sie im Nachfolgeprojekt QORA 2 geplant ist, der nächste Schritt hin zu einem besseren Verständnis der tatsächlich vorhandenen Fehlerquellen ist.

2.1.4 Erzeugung verschränkter Zustände (AP1.2)

Ziel dieser Arbeit war die Untersuchung verschränkter Zustände auf dem Quantencomputer ibmq_ehningen (AP1.2). Diese sollte zur Analyse der Qualität der Verschränkungsoperationen dienen und eine Abschätzung der Fehler ermöglichen. Hierfür wurde insbesondere die Erzeugung verschränkter sog. Greenberger-Horne-Zeilinger-Zustände (GHZ) untersucht. Neben bekannten Methoden zur Analyse der Qualität der Zustandserzeugung wurde nicht-adaptives messbasiertes Quanten-Computing implementiert. Dieses dient als eine weitere Nachweismethode für die Quanteneigenschaften der erzeugten Zustände – und damit einer möglichen Quanten-Überlegenheit.

Im ersten Teil von AP1.2 wurde die Erzeugung verschränkter Mehrteilchen-Zustände untersucht. Konkret wurden hierfür GHZ-Zustände mit bis zu 12 Qubits hergestellt. Da je nach Anzahl der Qubits eine Vielzahl an möglichen Konfigurationen existiert, wie solche Zustände mit ibmq_ehningen hergestellt werden können, wurden für jede Anzahl an Qubits (von 2 bis 12) 50 verschiedene Konfigurationen implementiert. Daraus wurde der Mittelwert über die Ergebnisse gebildet, um eine möglichst allgemeine Aussage zu erhalten.

Zur Analyse der jeweiligen Zustände wurde eine Methode angewandt, die die Abschätzung der Qualität der Zustände ermöglicht. Dabei war es besonders wichtig, eine effiziente Methode auszuwählen. Aus diesem Grund wurde die Methode „Multiple Quantum Coherences“ (MQC) [nach PRA **101**, 032343 (2020)] verwendet, welche im Gegensatz zur herkömmlichen Quantum State Tomography (QST) linear mit der Anzahl der Qubits skaliert und auf ibmq_ehningen übertragen werden kann. Zur Bestimmung der Qualität des jeweiligen Zustandes wurde jeweils eine obere und untere Schranke für die Fideliy (Güte) bestimmt und schließlich der arithmetische Mittelwert aller abgeschätzten Fidelities berechnet, siehe Abb. 1.8. Es wurden für jede Anzahl an Qubits jeweils zwei unterschiedliche Algorithmen zur Erstellung der Zustände angewandt (siehe Abb. 1.9). Die Fidelities erreichen Werte von 0.98 für zwei Qubits und fallen mit höherer Qubit-Anzahl ab, da hier die imperfekten Gatter des Quantencomputers zunehmend relevanter werden. Bei Zuständen mit mehr als 12 Qubits liegen die erreichten Fidelities unter 50%. Dies ist ein Indikator dafür, dass die zugehörigen GHZ-Zustände in dieser Größe für Quantenanwendungen nicht mehr vorteilhaft sind.

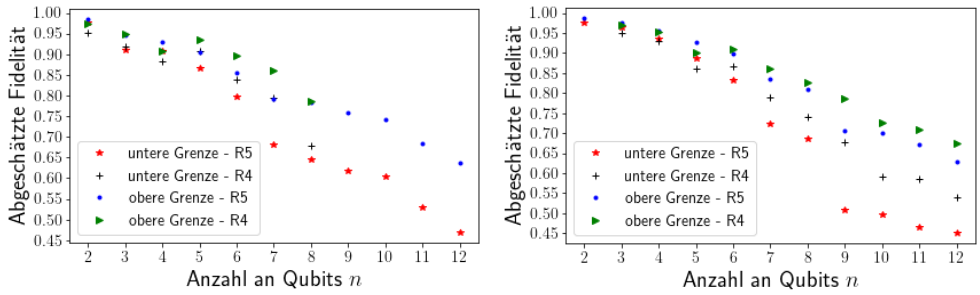


Abbildung 1.8: Durchschnittliche GHZ-Zustands-Fidelities (obere und untere Grenze) für verschiedene Anzahl an Qubits. Die linke und rechte Seite unterscheiden sich im Algorithmus zur Erstellung der Zustände (links: Optimierung nach „readout assignment error“ und rechts: Optimierung nach T_1-Zeit).

Die Untersuchungen wurden sowohl mit dem Prozessor Falcon R4 und Falcon R5 durchgeführt, da der Austausch der Prozessoren im Berichtszeitraum erfolgte. Im Rahmen unserer Untersuchung zeigt Falcon R5 keine verbesserten Ergebnisse im Vergleich zu Falcon R4.

Im zweiten Teil des Arbeitspakets AP1.2 wurde die aus der Quantenoptik bekannte Methode des nicht-adaptiven messbasierten Quanten-Computing (NMQC) implementiert. Diese Methode eignet sich besonders gut dazu, die Erzeugung von GHZ-Zuständen und deren Vielteilchenverschränkung zu testen, da GHZ-Zustände aufgrund der maximalen Quantenkorrelationen zwischen allen Zustandsqubits die bestmögliche Ressource für NMQC bilden. Die Idee von NMQC ist es, mithilfe der Messung eines verschränkten Zustands nichtlineare Boolesche Funktionen zu berechnen. Da hierbei die erfolgreiche Berechnung in direktem Zusammenhang zur Verletzung einer Bellschen Ungleichung steht, kann über den Test der Ungleichung gleichzeitig die Güte des verschränkten Zustands und der Berechnung überprüft werden. Die Mittelung dieser Tests über den gesamten Quantencomputer, also alle Qubitkonfigurationen, die einen GHZ-Zustand erzeugen können, ermöglicht es damit, das quantenmechanische Verhalten des Quantencomputers zu testen und zu charakterisieren.

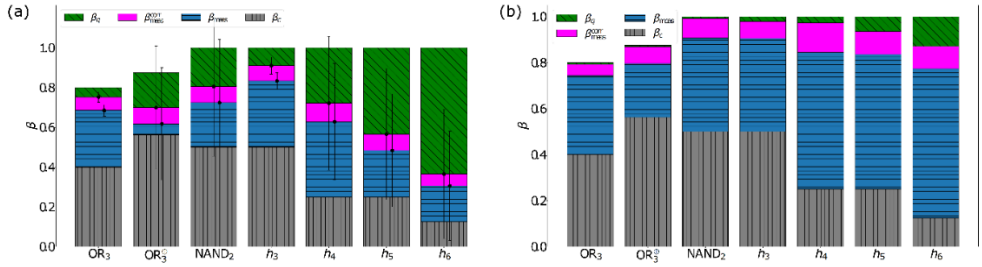


Abbildung 1.9: a) Durchschnittlicher gemessener Wert β_{meas} ($\beta_{\mathrm{meas}}^{\mathrm{corr}}$) ohne (mit) Fehlermitigation für die verschiedenen getesteten Vielteilchen-Bell-Ungleichungen. Während β_c den maximalen Wert darstellt, der mithilfe von klassischen Korrelation erreicht werden kann, stellt β_q den maximalen quantenmechanischen Wert dar. OR_3, $\mathrm{OR}_3^{\oplus}$, NAND_2, h_3 beschreiben Bell-Ungleichungen, die 4-Qubit-GHZ-Zustände testen. h_i beschreibt eine Bell-Ungleichung, die (i+1)-Qubit-GHZ-Zustände testet. Zwar sind β_{meas} bzw. $\beta_{\mathrm{meas}}^{\mathrm{corr}}$ in allen Fällen größer als der klassische Grenzwert, die großen Standardabweichungen lassen eine Verletzung allerdings nur bis mit zu 5 Qubits zu. b) Beste Werte β_{meas} ($\beta_{\mathrm{meas}}^{\mathrm{corr}}$), die für eine einzelne Qubit-Konfiguration gemessen wurden. Diese Werte sind wesentlich größer als die gemittelten Werte.

Im Gegensatz zur reinen Analyse der Fidelity der erzeugten GHZ-Zustände hat sich bei der Durchführung und Auswertung von NMQC gezeigt, dass Vielteilchen-Bell-Ungleichungen statistisch haltbar nur mit bis zu 5 Qubits verletzt werden konnten (siehe Abb. 1.9 a). Zwar gibt es einzelne Qubitkonfigurationen, welche sehr gute Werte liefern (siehe Abb. 1.9 b), diese werden jedoch wiederum durch Qubitkonfigurationen, welche lediglich eine gerine oder keine Verletzung aufzeigten, ausgeglichen. Um die gemessen Werte zu verbessern wurden Fehlermitigationsmethoden angewandt.

Insgesamt konnte mithilfe von GHZ-Zuständen und der Ausführung von NMQC auf ibmq_ehningen ein Vorteil von quantenmechanischen gegenüber klassischen Ressourcen gezeigt werden. Die Verletzung von Vielteilchen-Bell-Ungleichungen zeigt hierbei, dass der Vorteil aufgrund von Nichtklassizität, konkret Nichtlokalität, entsteht. Die Methode hat sich als geeignet erwiesen, das quantenmechanische Verhalten von Quantencomputern zu klassifizieren und kann in Zukunft auch auf andere Plattformen übertragen werden.

Als Highlight lässt sich der gemeinsame Gewinn von IBMs Open Science Prize durch Herrn Dr. Daniel Bhatti, Herrn Sebastian Brandhofer und Frau Jelena Mackeprang (Universität Stuttgart) anführen. Dabei wurde ein spezieller verschränkter Quantenzustand mit einer deutlich verbesserten Fidelity erzeugt.

2.1.5 Adaptivität des QAOA-Algorithmus gegenüber zeitlich korrelierten Fehlern (AP1.4)

Im AP1.4 wurde zunächst ein Modell für zeitlich korrelierte Fehler entwickelt, das wegen seiner Einfachheit zur Analyse der Resilienz von QAOA geeignet ist, jedoch physikalisch realistisch und aussagekräftig ist. Mithilfe dieses Modells und numerischer Simulationen haben wir detaillierte Erkenntnisse zur Fehlerresilienz gewonnen (Meilenstein M4), die künftig zu einer Steigerung der Fehlerresilienz variationeller Algorithmen wie z.B. QAOA führen könnten.

In dem von uns entwickelten Fehlermodel (siehe Abb. 1.10) befindet sich in der Nähe jedes Qubits ein Fluktuator, der nur mit diesem Qubit eine Wechselwirkung hat. (In supraleitenden Quantenrechnern, wie dem IBM Quantum System One, können diese Fluktuatoren durch Quasiteilchen auf den supraleitenden Kontakten realisiert sein.) Ein Fluktuator kennt zwei mögliche Zustände: angeregt oder nicht angeregt. In jedem Schritt eines Quantenschaltkreises tritt eine Wechselwirkung zwischen Fluktuator und Qubit auf. Falls der Fluktuator nicht angeregt ist, passiert kein Fehler auf dem Qubit. Falls der Fluktuator angeregt ist, passiert ein Bit- und Phasenflip-Fehler (Pauli-Y-Fehler). Zeitlich korrelierte Fehler entstehen dann dadurch, dass der Fluktuator eine innere Zeitentwicklung besitzt, die durch eine Anregungswahrscheinlichkeit p_F und Korrelationszeit τ gekennzeichnet wird. Dieses Model ist imstande, völlig unkorrelierte Fehler ($\tau = 0$), völlig korrelierte Fehler ($\tau \to \infty$), sowie auch teilweise korrelierte Fehler ($0 < \tau < \infty$) darzustellen.

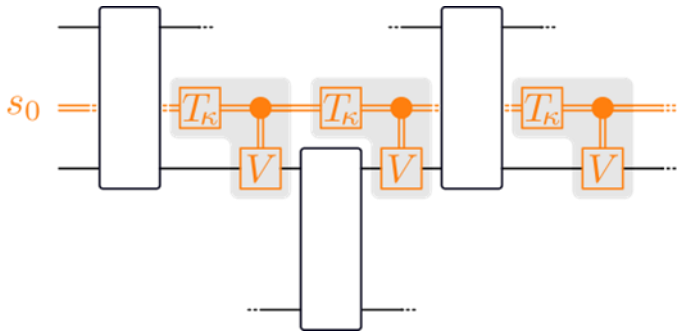


Abbildung 1.10: Das Modell der zeitlich korrelierten Fehler. Jedes Qubit interagiert mit einem unabhängigen klassischen Fluktuator (orange, es wird nur ein Fluktuator dargestellt). Der Fluktuator wird zu Beginn mit einer Wahrscheinlichkeit p_F angeregt, wie es vom klassischen Ensemble s_0 beschrieben wird. Die Operation T_κ setzt den Fluktuator mit einer Wahrscheinlichkeit von $1-\kappa$ zurück auf s_0. Der unitäre Fehleroperator $V = Y$ wird auf das Qubit angewendet, wenn der Fluktuator angeregt ist. Dies führt zu einem Fehlermodell mit unabhängiger zeitlich-lokaler Fehlerwahrscheinlichkeit p_F und Korrelationszeit $0 \leq \tau = -1/\ln\kappa < \infty$ (in Einheiten der Gate-Zeit). Das vollständige Fehlermodel wird erhalten durch Wiederholung des schattierten Bereichs.

Zur Analyse des Einflusses des Fehlermodells wurde die Abhängigkeit der „Approximation Ratio" (AR) als Funktion der Fehlerwahrscheinlichkeit p_F für 16 zufällige aber repräsentative QUBO-Probleminstanzen untersucht. Das AR ist ein Maß für die Qualität der Lösung eines Optimierungsverfahrens wie QAOA, definiert über das Verhältnis zwischen dem gefundenen Wert der Kostenfunktion und dem globalen, optimalen Wert der Kostenfunktion. Eine quadratische Abhängigkeit, AR(p_F)~1-p_F^2, würde als deutliches Merkmal der Fehlerresilienz gelten, weil in diesem Fall eine kleine Fehlerwahrscheinlichkeit p_F einer vernachlässigbaren Verringerung des AR zur Folge hat. Jedoch wurde eine lineare Abhängigkeit, AR(p_F)~1-p_F, beobachtet, was zunächst auf eine Abwesenheit der Fehlerresilienz hinzuweisen scheint. Es wurde jedoch nachgewiesen, dass QAOA nichtsdestotrotz eine andere Art von Fehlerresilienz besitzt. QAOA ist ein variationeller Optimierungsalgorithmus, der zur Lösung die Parameter eines

Lösungsansatzes so optimiert, dass das AR des Lösungsansatzes möglichst hoch wird. Im Fall von Rauschen, welches Fehler während der Berechnung der Kostenfunktion zufolge hat, wurde gezeigt, dass sich die Parameter durch die Optimierung an das Rauschen anpassen und so ein höheres AR erreicht werden kann. Die Parameter sind also „rauschbewusst".

Um den Vergleich zu nicht variationellen Algorithmen zu ermöglichen, wurde ein „rauschunbewusstes" AR definiert. Um dieses zu berechnen, wird in den Simulationen zunächst das Rauschen ausgeschaltet und die optimalen Parameter bestimmt. Die so erhaltenen optimalen Parameter sind also nicht an das Rauschen angepasst. Anschließend wird das Rauschen wieder eingeschaltet und das AR aus dem Lösungsansatz mit den rauschunbewussten Parametern berechnet. Dies ist das rauschunbewusste AR. Der Unterschied zwischen dem rauschbewussten und rauschunbewussten AR ist ein Maß für die Rauschadaptivität des QAOAs.

Für 16 zufällige aber typische Probleminstanzen auf 6 Qubits haben wir eine nicht vernachlässigbare Rauschadaptivität nachgewiesen. Die Rauschadaptivität tritt jedoch erst über einem gewissen, von der Probleminstanz abhängigen Schwellenwert von p_F auf. Die Daten einer Probleminstanz werden in Abb. 1.11 gezeigt.

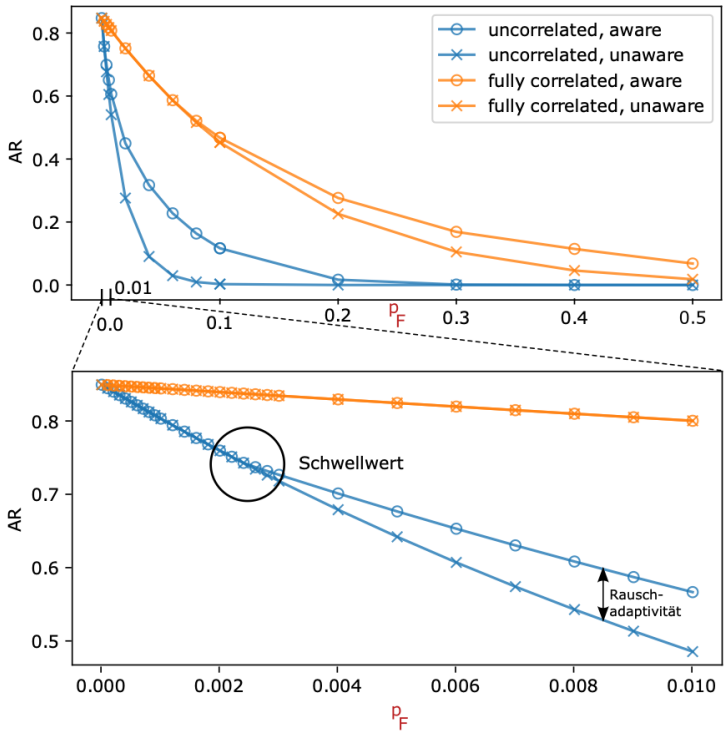


Abbildung 1.11: Approximation Ratio (AR) als Funktion der Fehlerstärke p_F, sowohl für die rauschbewusste als rauschunbewusste AR, im Fall von völlig korrelierten und völlig unkorrelierten Fehlern. Die Rauschadaptivität tritt erst überhalb eines Schwellwerts von p_F auf.

Darüber hinaus zeigen unsere Ergebnisse, dass eine große Korrelationszeit zu einer erhöhten AR führt, sowohl für die rauschbewusste als rauschunbewusste AR. Dies spricht deutlich für die praktische Ausführbarkeit von QAOAs im Vergleich zu konventionellen, fehlerkorrigierten, nicht-variationellen Algorithmen, da letztere bei steigenden Korrelationszeiten einen steigenden Mehraufwand erfordern. Diese Ergebnisse wurden bei Physical Review A eingereicht und sind dort mittlerweile für die Veröffentlichung als „Editor's suggestion" akzeptiert [4].

2.1.6 Funktionaler Test von Quantenchips (AP1.6)

Im Arbeitspaket AP1.6 wurden funktionale Tests von 1-Qubit-Gattern auf dem 27-Qubit-System ibmq_ehningen durchgeführt. Ziel war es, mit Hilfe der in [Milazzo et al.] vorgestellten Bayesschen adaptiven Techniken schnellstmöglich zu testen, ob die Gatter wie geplant funktionieren. Der Schwerpunkt des Testverfahrens liegt nicht auf einer vollständigen Charakterisierung der Fehler, sondern auf einer möglichst schnellen Beurteilung der Funktionalität eines gegebenen Quantenprozessors. Im Zuge der adaptiven Bayesschen Schätzung wird mit einer gleichmäßigen A-priori-Verteilung des zu testenden Parameters begonnen. Darauffolgende Messungen werden dahingehend optimiert, maximalen Informationszuwachs zu erhalten, um so keine Zeit mit suboptimalen Messungen zu verschwenden. Mit den Messergebnissen wird die A-posteriori-Verteilung des Parameters aktualisiert.

Bei jedem Schritt werden die bereits gewonnenen Informationen über den Wert des Parameters verwendet, um über die beste Vorgehensweise bei den nächsten Messungen zu entscheiden. Für unsere Anwendung entspricht dies einer Entscheidung zwischen einer X oder Y Messung. Für die Implementierung haben wir den Funktionstest des Phasengatters, welches als virtuelle Rotation fehlerfrei von IBM implementiert werden kann, an jedem der 27 Qubits des Ehninger Geräts durchgeführt. Hierbei nutzen wir zwei verschiedene Strategien. Bei der ersten Strategie haben wir die erforderlichen Gatter nacheinander angewendet, was es ermöglicht, potenzielle Effekte von Cross-Talk zu vermeiden, aber gleichzeitig die Tiefe der Schaltungen und die Dauer des Testverfahrens erhöht. Die so erhaltenen experimentellen Ergebnisse sind zusammen mit simulierten Daten in Abbildung 1.13 dargestellt.

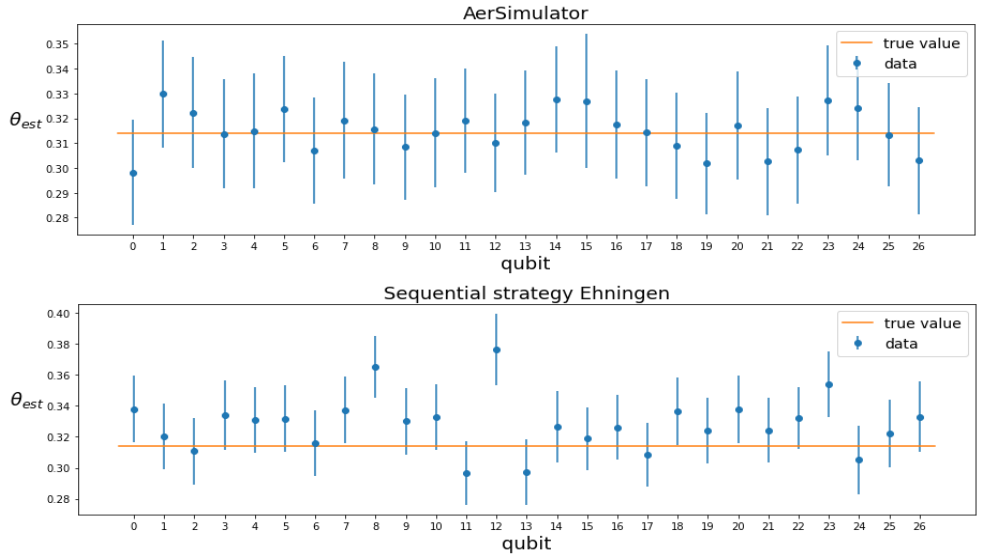


Abbildung 1.12: Vergleich zwischen den im Simulator und im realen Gerät erzielten Ergebnissen bei Verwendung einer sequenziellen Strategie. Die blauen Kreise stellen den Durchschnitt der A-posteriori-Verteilung nach $N = 30$ Bayesschen Aktualisierungen mit jeweils $n_{shots} = 300$ Shots dar. Der Wert der zu implementierenden Phase wurde auf $\theta = \pi/10$ festgelegt. Die Fehlerbalken stellen die Unsicherheit in der A-posteriori-Verteilung dar, quantifiziert durch zwei Standardabweichungen. Der tatsächliche Wert liegt im Simulator immer innerhalb von zwei Standardabweichungen. Auf dem realen Gerät hingegen liefern einige Qubits Ergebnisse, die deutlich von dem Wert entfernt sind, der ursprünglich implementiert werden sollte.

Bei der anderen Teststrategie wurden die Gatter alle parallel auf die 27 Qubits angewendet, was Schaltungen mit geringerer Tiefe ermöglichte, aber möglicherweise einige Cross-Talk Effekte während des Testverfahrens verursachte. Die mit der parallelen Strategie erzielten Ergebnisse sind in Abbildung 1.14 zu sehen. Deutlich wird, dass die geschätzte Phase teilweise stark von der eigentlich auf dem Gerät zu implementierenden

Phase abweicht. Dies könnte eine Konsequenz ungewünschter Effekte aufgrund der parallelen Anwendung sein.

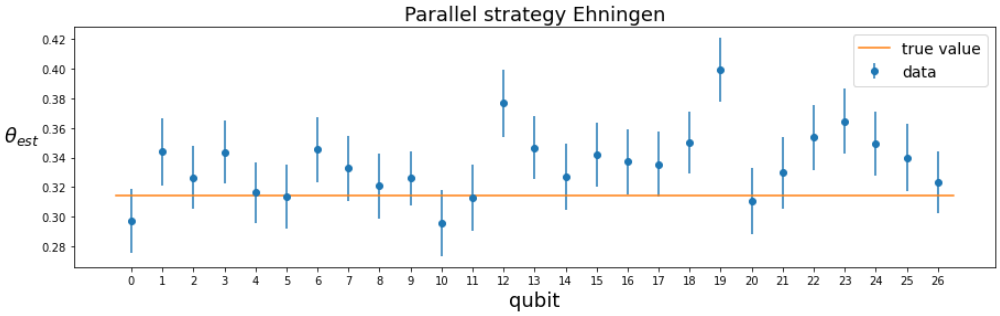


Abbildung 1.13: Ergebnisse, die auf dem realen Gerät mit einer parallelen Strategie nach $N = 30$ Bayessche Aktualisierungen mit jeweils $n_{shots} = 300$ Shots erhalten wurden. Die Abweichung vom wahren Wert ist größer als die Unsicherheit (zwei Standardabweichungen in der A-posteriori-Verteilung) für eine größere Anzahl von Qubits im Vergleich zu den Ergebnissen, die mit der sequentiellen Strategie erzielt wurden. Dies bestätigt, dass die parallele Anwendung der Gatter die Qualität Parameterabschätzung verschlechtert.

Sowohl bei der sequentiellen als auch bei der parallelen Strategie wurden die Ergebnisse durch $N = 30$ Bayessche Aktualisierungen erhalten. Die Messungen für jede Aktualisierung wurden mit der relativ geringen Anzahl von $n_{shots} = 300$ Shots erhalten. Informationen aus vorherigen Messungen wurden genutzt, um die A-posteriori-Verteilung des Parameters zu aktualisieren und deren Varianz zu verringern (siehe Abbidung 1.15). So konnten vorhandene Informationen aus vorherigen Messungen optimal genutzt und die Anzahl benötigter Shots entsprechend reduziert werden.

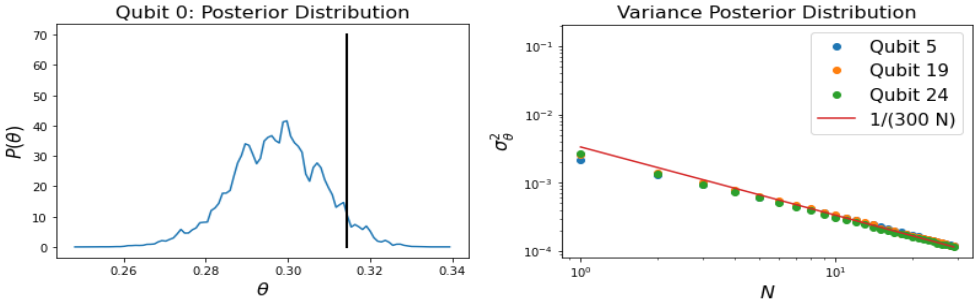


Abbildung 1.14 (Links): Beispiel der A-posteriori-Verteilung, die mit dem Adaptiven Bayesschen Funktionstest (parallele Strategie) nach $N = 30$ Aktualisierungen mit jeweils $n_{shots} = 300$ Shots erhalten wurde. (Rechts) Varianz der A-posteriori-Verteilung für den Parameter θ aufgetragen gegen die Anzahl N der Bayesschen Aktualisierungen. Zu sehen sind die Ergebnisse für Qubit 5, welches ordnungsgemäß funktioniert, und Qubit 19 und 24, welche nicht wie vorgesehen funktionieren.

2.2 AP 2: Fehlerreduzierung

Das Ziel des zweiten Arbeitspakets war die Erhöhung der Fehlerresilienz von Quantenschaltungen. Dabei haben sich sowohl in diesem Projekt als auch in der weltweiten *Scientific Community* die Transpilationsverfahren als das probate Mittel herauskristallisiert, um die Erfolgswahrscheinlichkeit einer Berechnung auf einem physikalischen Quantencomputer zu erhöhen. Sie schlagen ferner eine Brücke zu Fehlermitigationsverfahren, die einen Schwerpunkt des Folgeprojekts QORA II bilden. Neben den bereits im Zwischenbericht erörterten Publikationen [1] und [6] (letztere wurde nach der Abgabe des Zwischenberichts zur Veröffentlichung in der Zeitschrift *Quantum Information Processing* angenommen) haben Arbeiten der zweiten Projekthälfte zum Thema Transpilation zu Veröffentlichung [2] bei der *IEEE Quantum Computer Engineering* Konferenz und zu Preprints [7] und [8] geführt (siehe Kap. 2.2.1 und 2.2.3), die sich derzeit in Begutachtung befinden. Der Schwerpunkt der Darstellung in Kap. 2.2.1 liegt auf neueren Arbeiten aus der zweiten Projekthälfte; die in Kap. 2.2.2 berichteten Ergebnisse zu Simulationen mit Hilfe des verbesserten Fehlermodell aus Kap. 2.1.3 führten zu dem Preprint [9].

2.2.1 Transpilationsverfahren (AP2.2)

Bei der Transpilation handelt es sich um ein Verfahren, um eine Quantenschaltung auf einem gegebenen Quantenrechner lauffähig zu machen. So zeigt Abb. 2.1 die Topologie des Quantenrechners in Ehningen; zwei-Qubit-Operationen sind nur zwischen verbundenen Qubits möglich. Rechts in Abb. 2.1 ist eine 5-Qubit-Schaltung gezeigt, die auf Qubits 16, 19, 20, 22 und 25 des Ehninger Rechners transpiliert ist. Für diese sehr einfache Schaltung werden alle notwendigen Verbindungen unterstützt, im Allgemeinen müssen aber zusätzliche „Ancilla-Qubits“ verwendet werden. Außerdem enthält Abb. 2.1 Angaben über die aktuell beobachteten Fehlerraten; so weisen Qubit 15 und die Operation zwischen Qubits 2 und 3 zum betrachteten Zeitpunkt erhöhte Fehlerraten auf und sollten bei der Transpilation nach Möglichkeit gemieden werden.

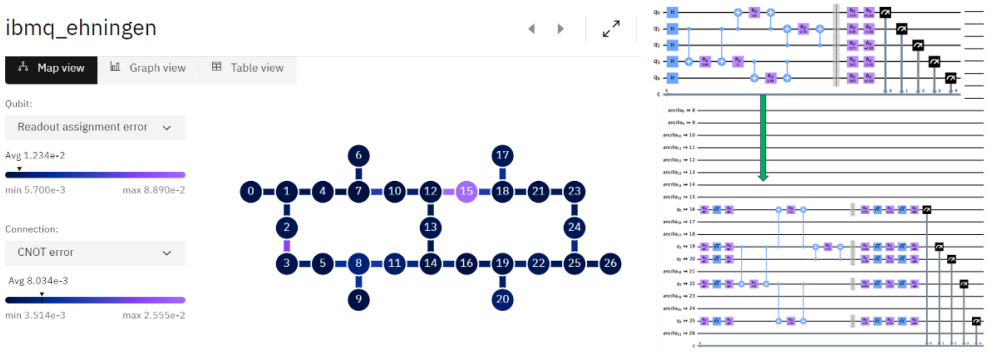


Abbildung 2.1: Links: Topologie des Quantenrechners in Ehningen und Angaben zu Fehlerraten. Rechts: Transpilation von einer Quantenschaltung mit 5 Qubits auf den Quantenrechner in Ehningen.

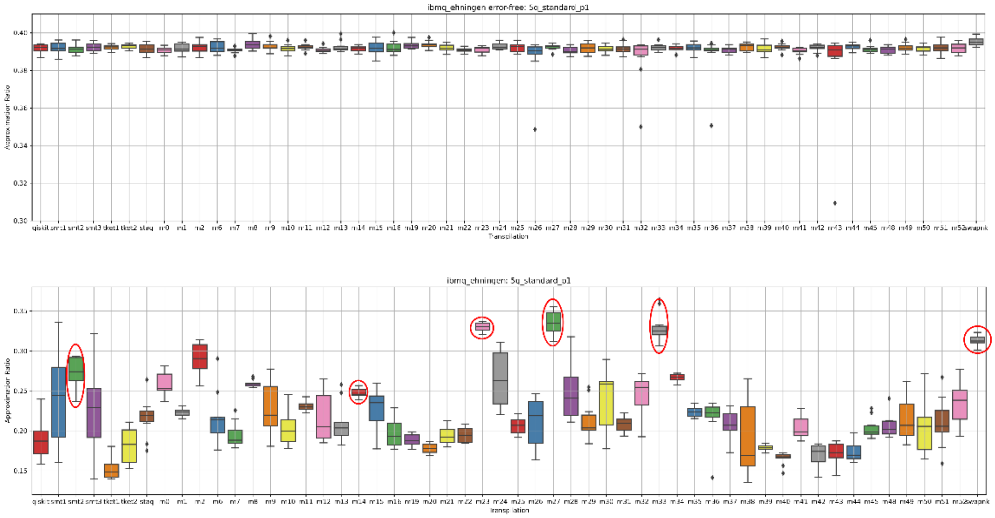


Abbildung 2.2: Vergleich verschiedener Transpilationsverfahren. Für jedes Verfahren wird die durch QAOA erreichte Approximation Ratio für eine fehlerfreie Simulation (oben) und für den Quantenrechner in Ehningen (unten) dargestellt. Die ausgewählten sechs Verfahren werden durch rote Kreise verdeutlicht und erzielen bessere Ergebnisse als das vom IBM im Softwarepaket „Qiskit" bereitgestelle Verfahren (dunkelroter Balken ganz links).

Im Bereich Transpilation haben wir eine umfassende Evaluierung des Stands der Technik durchgeführt und existierende Werkzeuge (tket, staq, Qiskit) eingesetzt. Daraufhin wurden eigene Formulierungen von Transpilationsfragestellungen basierend auf *Satisfiability Modulo Theory* (SMT) aufgestellt, die jetzt entweder anstelle oder zusammen mit existierenden Werkzeugen verwendet werden können. Bei SMT-Formulierungen lässt sich ein konkretes Optimierungsziel (z.B. Minimierung der Gatteranzahl, der Laufzeit oder der erwarteten Ausfallwahrscheinlichkeit) angeben, die dann beweisbar optimal erreicht wird. Transpilation lässt sich in das Problem der initialen Qubit-Zuweisung (*Initial Mapping*) und das *Routing*-Problem aufteilen. Wir haben insgesamt über 60 Kombinationen von Verfahren zur Lösung dieser Teilprobleme simulativ und durch Experimente auf dem Quantenrechner evaluiert (vgl. Abb. 2.2). Dabei haben sich sechs Varianten (sowohl mit als auch ohne SMT) herauskristallisiert, die durchgehend besonders gute Ergebnisse liefern und die nun vertieft untersucht werden. Außerdem wurde gemeinsam mit der Universität Konstanz der Einsatz von allgemeineren *Swap Networks* als Alternative zur Transpilation untersucht; diese sind für spezielle Topologien beweisbar asymptotisch optimal, die auf dem Quantenrechner in Ehningen allerdings nicht gegeben sind.

Calibration-Aware Transpilation

Das neuartige Konzept einer *Calibration-Aware Transpilation* wurde in [2] vorgeschlagen. Es basiert auf der experimentell auf dem IBM-Quantencomputer in Ehningen unterfütterten (Abb. 2.3) Beobachtung, dass sich die Fehlerraten nicht nur zwischen den einzelnen Qubits sondern auch auf demselben Qubits über die Zeit verändern. Somit ist eine gute Transpilationslösung nach Verstreichen einer gewissen Zeit nicht mehr optimal, weil sie auf veralteten Daten basiert; IBM führt stündliche Kalibrationen durch und stellt mindestens alle 24 Stunden aktuelle Fehlerraten bereit. Bei der *Calibration-Aware Transpilation* wird der aufwändige Transpilationsvorgang in drei Schritte unterteilt, wobei der komplexeste erste Schritt („Topology-Aware Pre-Transpilation") nur einmal durchgeführt wird. Bei relevanter Veränderung von Fehlerraten nach einer Kalibration muss nur noch der deutlich schnellere zweite Schritt („Noise-Aware Matching") mit aktuellen Fehlerraten wiederholt werden. Speziell für inkrementelle Algorithmen wie das

in QORA schwerpunktmäßig betrachtete QAOA, die eine Folge von sehr ähnlichen aber nicht identischen *Ansatz*-Schaltungen ausführen, wurde noch ein dritter Schritt („Decomposition and Optimization“) hinzugefügt, der inkrementelle Anpassungen vornimmt.

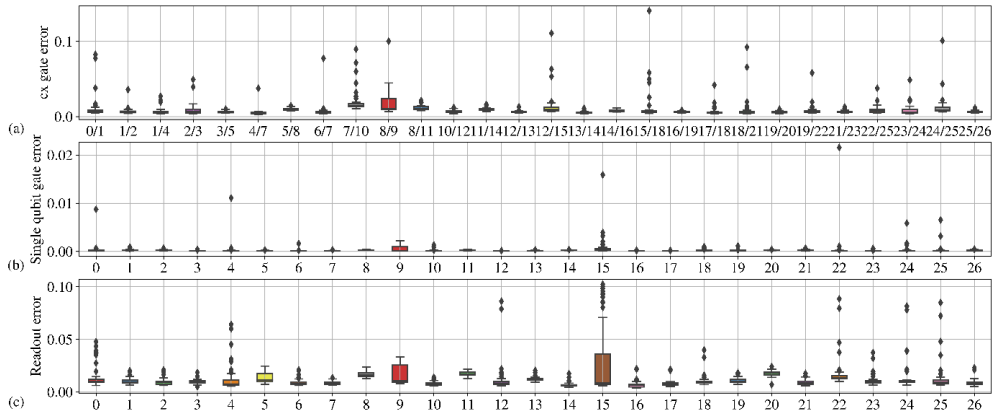


Abbildung 2.3: Entwicklung der Fehlerraten auf den einzelnen Qubits bzw. Qubit-Verbindungen des IBM-Quantenrechners in Ehningen über den Zeitraum von 39 Tagen.

Wie die Ergebnisse aus [2] in Abb. 2.4 zeigen, demonstriert *Calibration-Aware Transpilation* eine sehr kompetitive Performanz: QAOA-Ansatzschaltungen sind verglichen mit von anderen Verfahren transpilierten (und ansonsten gleichen) Schaltungen weniger von Fehlern betroffen und führen so zu einem besseren *Approximation Ratio* des Gesamtalgorithmus‘. Wir haben experimentell gezeigt, dass durch die Unterteilung in drei Schritte nur vernachlässigbare Laufzeiteinbußen verglichen mit monolithischen Transpilationsverfahren vergleichbarer Qualität in Kauf zu nehmen sind. Prinzipiell sind die Schritte 2 und 3 so schnell, dass sie bei einer Abfolge von Ansatz-Schaltungen direkt zwischen die Abarbeitung der Schaltungen durchgeführt werden könnten und so das *Queing* einer jeden neu transpilierten vermeiden. Während die aktuellen IBM-Systeme ein solches *Feature* nicht unterstützen, zeigen unsere Beispielrechnungen, dass sich damit Beschleunigungen von ca. Faktor 10 realisieren lassen.

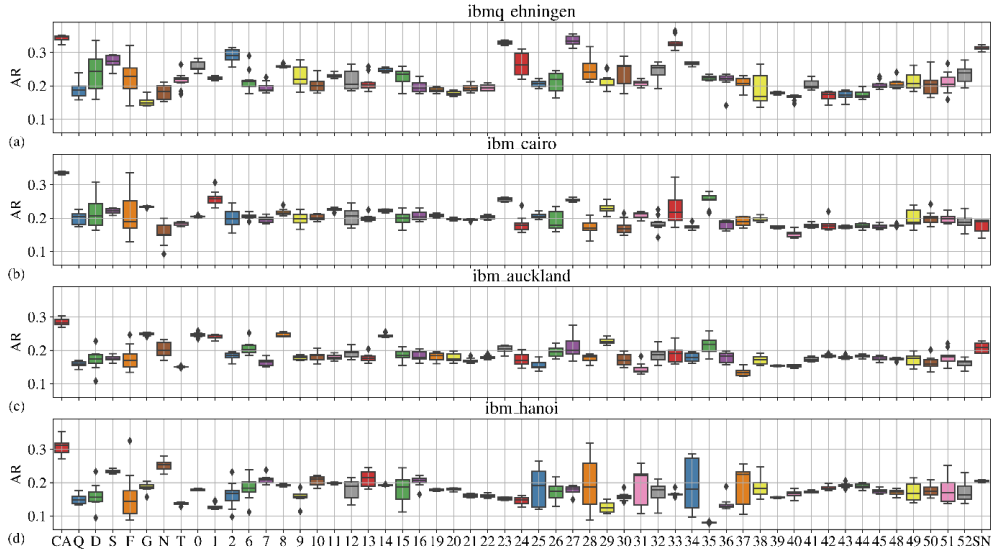


Abbildung 2.4: Approximation Ratio, das von unserem Verfahren („CA“, ganz links) erreicht wurde, im Vergleich mit anderen bekannten Transpilationsverfahren auf drei IBM-Quantenrechnern.

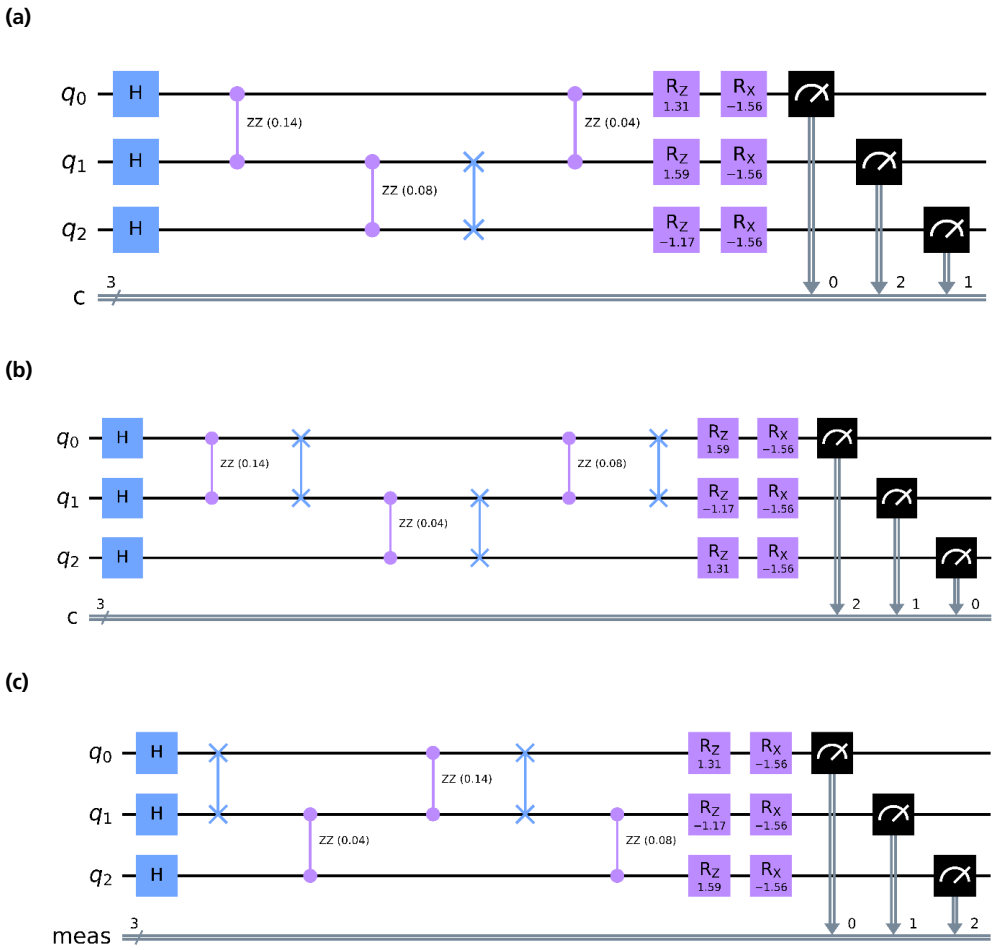
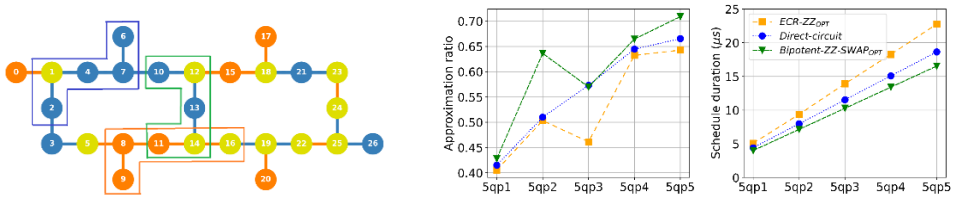


Abbildung 2.5: Ergebnis von Algorithm-Aware Qubit Mapping (a) für eine QAOA-Ansatzschaltung mit 3 Qubits verglichen mit SWAP-Network (b) und dem Transpilationsergebnis von QisKit (c) für die gleiche Schaltung

Während Transpilationsverfahren grundsätzlich generisch sind, also beliebige Quantenschaltungen mit oder ohne Struktur verarbeiten können, wurde in [7] ein Konzept erarbeitet, welches genau das Strukturwissen ausnutzt und so bessere Ergebnisse für spezielle Schaltungsklassen liefern kann. Unter dem Namen *Algorithm-Aware Qubit Mapping* wurde die Leistungsfähigkeit des Konzepts speziell für QAOA-Ansatzschaltungen auf bestimmten Qubit-Topologien demonstriert. So wurde unter anderem gezeigt, dass die *SWAP-Networks* entgegen der breit geteilten Meinung nicht die beweisbar kompaktesten Implementierungen auf linearen Topologien sind. Die berechneten Schaltungen haben eine (zum Teil deutlich) geringere Tiefe und Gatteranzahl als die Ergebnisse anderer, generischer Transpilationsverfahren. Dies wird in Abb. 2.5 exemplarisch bereits für die sehr kleine 3-Qubit-Ansatzschaltung demonstriert: Unser Verfahren erreicht die Tiefe von 18 (die Messoperationen nicht mit eingerechnet), die Schaltung mit dem *SWAP-Network* hat die Tiefe von 20 und die von QisKit berechnete Schaltung hat die Tiefe von 21. Experimente für QAOA mit begrenzter Anzahl von Qubits auf mehreren IBM-Quantencomputern haben auch hier eine bessere Performanz von QAOA mit der vorgeschlagenen Methode demonstriert.

Am Projektende wurde noch eine Fragestellung betrachtet, die zum Beantragungszeitpunkt nicht absehbar war. Der während der Projektlaufzeit neu eingeführte *Falcon*-Chipsatz unterstützt verbesserte CX-Gatter (sog. *Direct-CX*), bietet sie aber nicht auf allen Qubits. Die anderen Qubits müssen auf das konventionelle *ECR-CX*-Gatter zurückgreifen, das einerseits eine längere Dauer und eine schlechtere Fidelity

aufweist, andererseits aber bestimmte Optimierungen auf der Puls-Ebene ermöglicht, welche für *Direct-CX* zumindest derzeit nicht zur Verfügung stehen. Der Frage, wie eine Transpilation auf einer solchen Architektur, die wir „bipotent" nennen, durchzuführen ist, widmet sich die Arbeit [8]. Dabei werden sowohl die beiden Gattertypen durch physikalische Messungen einzeln untersucht als auch ihr Einsatz in QAOA-Ansatzschaltungen untersucht; für letztere werden verschiedene Architekturen mit unterschiedlichen Gattertypen konstruiert. Etwas überraschend zeigen unsere Ergebnisse, dass die positiven Effekte der Optimierungen auf Pulsebene zumindest derzeit die Vorteile des neuen Gatters überwiegen. Dies wird in Abb. 2.6 exemplarisch gezeigt.



(a) (b) (c)

Abbildung 2.6: Drei alternative *Mappings* auf der bipotenten IBM-Architektur (a). Blaue Qubits unterstützen neuartige *Direct-CX*-Gatter, orangene Qubits unterstützen ausschließlich die älteren *ECR-CX*-Gatter, grüne Qubits unterstützen beide. Das *Mapping Direct-circuit* verwendet ausschließlich *Direct-CX* Gatter und besteht aus Qubits 2, 1, 4, 7 und 6; *ECR-ZZ*$_{OPT}$ verwendet ausschließlich *ECR-CX*-Gatter und einige Optimierungen auf Pulsebene und besteht aus Qubits 9, 8, 11, 14 und 16; *Bipotent-ZZ-SWAP*$_{OPT}$ verwendet beide Gattertypen und alle verfügbaren Pulsebenen-Optimierungen; es besteht aus Qubits 10, 12, 13, 14 und 11. In (b) und (c) werden das *Approximation Ratio* und die Ausführungszeit für verschiedene Werte der QAOA-Tiefe *p* dargestellt; man sieht, dass das bipotente Mapping die bessere Lösungqualität liefert.

2.2.2 Simulationen mit Fehlermodellen aus AP1.5 (AP2.4)

Bei den Fehlermodellen haben sich die bisher vorgestellten Arbeiten auf die Standard-Modelle sowie die von IBM bereitgestellten rechnerspezifische Fehlerbeschreibungen konzentriert. Nachdem wir erste Ergebnisse physikalischer Messungen zur Verfügung hatten, arbeiteten wir an der Verfeinerung dieser Modelle insbesondere im Kontext der Transpilation. Einen Fokus der Untersuchungen stellen sog. *Crosstalk*-Effekte dar, also Beeinträchtigungen von Operationen durch eigentlich unabhängige Berechnungen, die auf benachbarten Qubits ablaufen (siehe auch Kapitel 2.1.3). Letztere wurden in AP1.5 im Detail untersucht und ein geeignetes, erweitertes Fehlermodell, welches durch Crosstalk ausgelöste Korrelationseffekte berücksichtigt, erstellt.

Um dieses erweiterte Crosstalk-empfindliche Fehlermodell zu testen, verwenden wir als Beispiel den aus 8 Qubit bestehenden „Hadamard-Leiter-Schaltkreis". Dieser ist in Abb. 2.7 zusammen mit der idealen Wahrscheinlichkeitsverteilung dargestellt, welche durch eine Messung der Qubits in der *Z*-Basis zustande kommt. Bevor wir nun den Hadamard-Leiter-Schaltkreis auf ibmq_ehningen ausführen und die Ergebnisse mit dem auf unserem Fehlermodell basierenden Simulator vergleichen, bestimmen wir zunächst alle 132 möglichen Qubit-Layouts der 27-Qubit-Architektur, auf welchen wir den 8-Qubit-Schaltkreis abbilden können. Um die Reproduzierbarkeit unserer Resultate zu überprüfen, wiederholen wir die Messung aller Layouts zweimal nacheinander und führen die Charakterisierung (siehe Abb. 1.6) dazwischen durch. Auf diese Weise können wir Fluktuationen identifizieren, welche sich schon über kurze Zeiträume auswirken. Außerdem wiederholen wir das gesamte Prozedere an verschiedenen Tagen, um mögliche Langzeiteffekte zu identifizieren. Des Weiteren ist zu bemerken, dass die Simulationen mittels unseres Fehlermodells ein gemitteltes Szenario darstellen, da es auf

gemessenen *mittleren* Fehlerraten beruht. Um die Simulationen trotzdem bestmöglich mit den Hardware-Experimenten zu vergleichen, haben wir im letzteren Fall die Schaltkreise mit zusätzlichen zufälligen Pauli-Gattern versehen. Letztere modifizieren das Hardware-Rauschen und transformieren es im Mittel in sog. Pauli-Rauschen, lassen den rauschfreien Schaltkreis jedoch unverändert. In der Praxis führt diese Methode zu einer höheren Anzahl an Schaltkreisen, welche aber mit einer niedrigeren Anzahl an „Shots" gemessen werden können, um die Gesamtzahl der „Shots" konstant zu halten. Bei allen Hardware-Experimenten haben wir alle sonstigen Fehlervermeidungs- oder Fehlerreduzierungs-Optionen des IBMQ Systems ausgeschaltet.

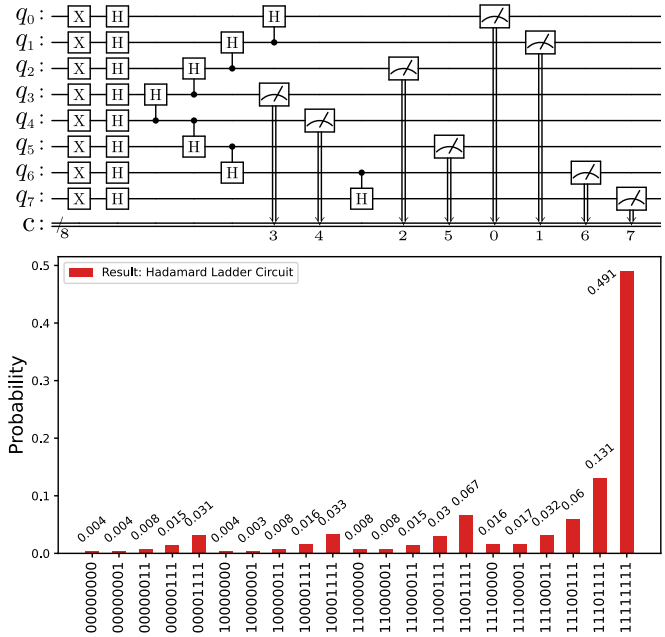


Abbildung 2.7: Darstellung des 8-Qubit-Hadamard-Leiter-Schaltkreises (oben) und der Wahrscheinlichkeitsverteilung (unten), welche eine Messung im fehlerfreien Fall ergibt. Dieser Schaltkreis dient zum Test unseres Crosstalk-sensitiven Fehlermodells in Abb. 8.

In Abb. 2.8 sind die Resultate der genannten Experimente dargestellt, wobei wir die gemessenen Wahrscheinlichkeitsverteilungen der einzelnen Hardware-Experimente und Simulationen jeweils anhand der sog. Hellinger-Güte vergleichen. Zusammenfassend finden wir, dass die Simulationen mit Hilfe des Crosstalk-sensitiven Fehlermodells die Hardware-Daten besser beschreiben als die Simulationen mittels des Standard-Fehlermodells, welche alleine auf den IBM-Kalibrierungsdaten beruhen. Dies wird auch durch die mittleren quadratischen Abweichungen ρ zwischen simulierter und gemessener Güte, welche wir für jeden Plot berechnen, bestätigt. In der Zeile (c) von Abb. 2.8 vergleichen wir außerdem die Güten der beiden hintereinander ausgeführten Hardware-Runs, welche zeigen, dass kurzzeitige Fluktuationen auftreten können, deren Stärke jedoch vom genauen Tag der Ausführung abhängen. Wenn wir uns also zusätzlich auf die Schaltkreise beschränken, bei denen die Abweichungen zwischen Run 1 und 2 nicht mehr als 2% betragen, sollten die entsprechenden Simulationen noch bessere Resultate liefern. Dies wird auch wieder durch die entsprechenden mittleren quadratischen Abweichungen (jeweils in Klammern) bestätigt.

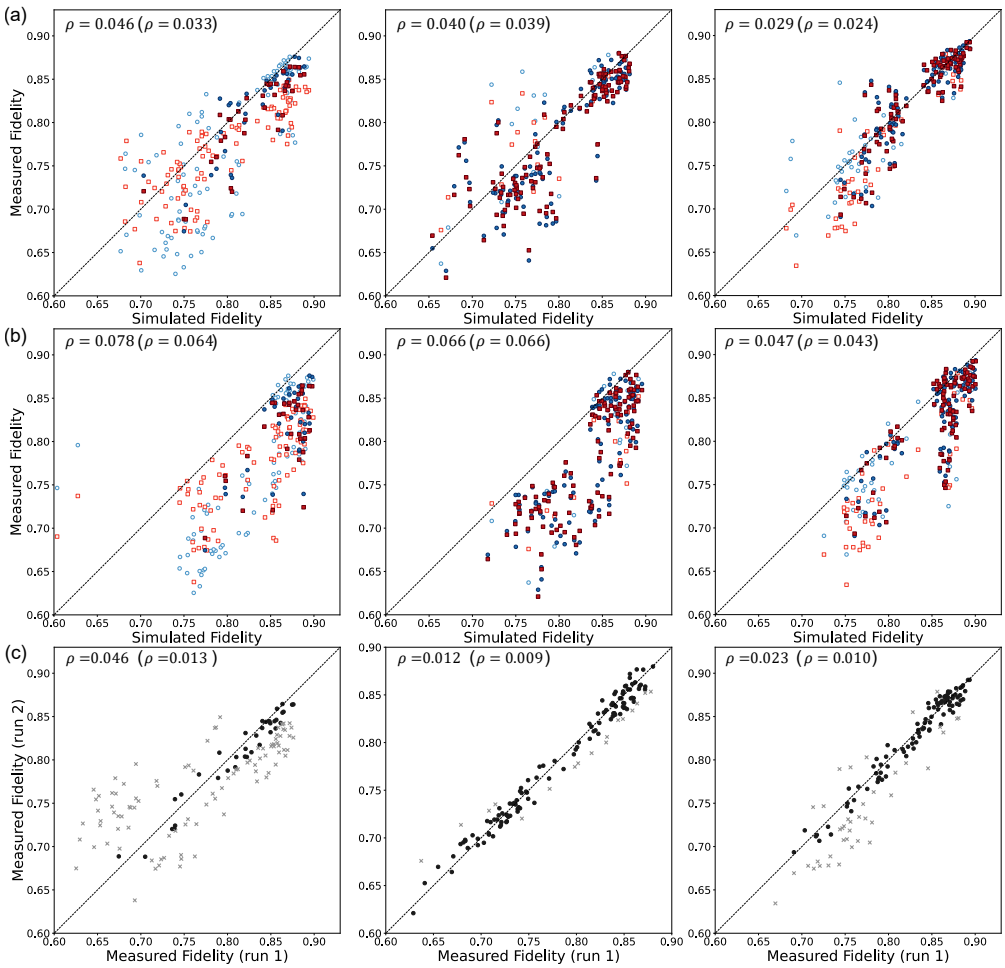


Abbildung 2.8: Vergleich der auf ibmq_ehningen (Falcon r5.11 Prozessor) gemessenen Hellinger-Güten des 8-Qubit-Hadamard-Leiter-Schaltkreises mit den simulierten Experimenten am 20. (linke Spalte), 23. (mittlere Spalte) und 24. (rechte Spalte) Januar 2023. Jede Spalte enthält die Resultate von zwei hintereinander ausgeführten Hardware-„Runs" aller 132 Qubit-Layouts (jeweils dargestellt durch Kreise und Quadrate). (a) Korrelationen zwischen den gemessenen und mit Hilfe unseres Crosstalk-sensitiven Fehlermodells simulierten Güten. (b) Korrelationen zwischen den gemessenen und mit Hilfe des standardmäßigen Fehlermodells simulierten Güten. (c) Korrelationen zwischen den Güten der beiden nacheinander gemessenen Hardware-Experimente. Graue Kreuze zeigen die Güten derjenigen Qubit-Layouts, bei denen eine Abweichung von mehr als 2% zwischen Run 1 und Run 2 auftritt. Diese Layouts sind in den Plots (a) und (b) in heller Farbe und durch leere Symbole dargestellt.

2.2.3 Ausnutzung resilienter Teilstrukturen (AP2.6)

Während die oben beschriebenen Transpilationsverfahren traditionelle Fehler auf Qubits und Qubit-Verbindungen annehmen, wurden auch einige Arbeiten zur Transpilation unter *Crosstalk*-Fehlern durchgeführt. Die Idee dabei ist es, Teilstrukturen des Quantenprozessors zu identifizieren, auf denen die tatsächlich beobachteten Fehler größer sein können als in den Kalibrierungsdaten berichtet. Dieses Verhalten ist in vielen Fällen auf Crosstalk zwischen benachbarten Qubits zurückzuführen (siehe Kap. 2.1.2). Informationen über die Stärke dieser korrelierten Fehler können also ausgenutzt werden um fehlerresilienten Teilstrukturen zu identifizieren und die betrachtete Quantenschaltkreise darauf auszuführen.

Bei den Arbeiten zur Charakterisierung von korrelierten Fehlern (AP1) wurden für die Transpilation im Wesentlichen zwei Metriken erhoben: der Zwei-Qubit Gatterfehler, wenn ein benachbartes Qubit aktiv ist und der Einzel-Qubit Gatterfehler, wenn ein Zwei-Qubit-Gatter auf einem benachbarten Qubitpaar ausgeführt wird. Diese beiden Metriken wurden zur Erreichung von Meilenstein M5 in ein Transpilationsverfahren eingebettet, das eine optimierte Auswahl der vorhandenen Quantenrechner-Qubits für die Berechnung einer gegeben Schaltung bestimmt.

Abbildung 2.9 visualisiert das entwickelte Transpilationsverfahren für korrelierte Fehler. Zunächst werden die Anforderungen an Konnektivität und Anzahl der Qubits für die zu transpilierende Zielschaltung (links unten) bestimmt. In diesem Beispiel werden 4 Qubits benötigt, wobei ein Qubit direkt mit den drei anderen Qubits interagieren können muss. In dem Konnektivitätsgraphen von ibmq_ehningen (rechts oben) ergeben sich dadurch 8 Platzierungsmöglichkeiten, aus denen die Platzierung mit der höchsten Erfolgswahrscheinlichkeit bzw. mit dem geringsten erwarteten Fehler bestimmt wird. Die Platzierungsmöglichkeiten einer Zielschaltung auf einem gegebenen Quantenrechner werden durch das Lösen eines Subgraph-Isomorphie-Problems bestimmt [Paul D. Nation and Matthew Treinish, PRX Quantum **4**, 010327 (2023)].

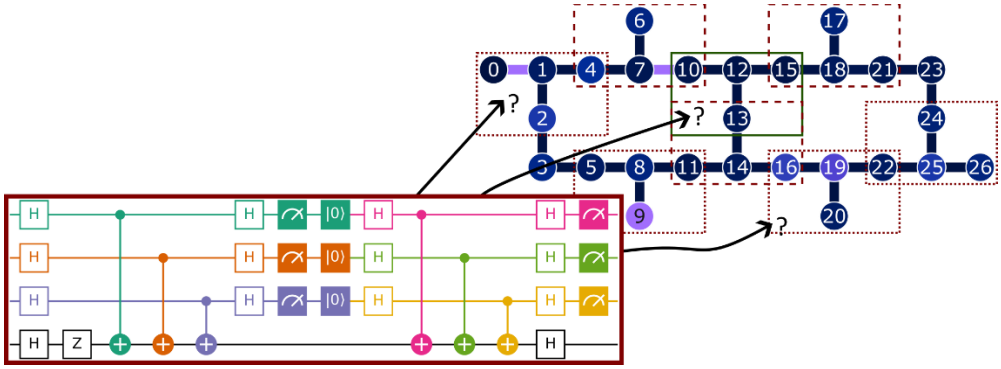


Abbildung 2.9: Transpilationsverfahren für korrelierte Fehler

Die Platzierungsmöglichkeiten werden bezüglich des erwarteten Fehlers evaluiert. Dafür wird eine Kostenfunktion definiert, die alle berücksichtigen Fehlerquellen einbezieht. Im Fall der korrelierten Fehler werden neben erwarteten Quantengatter- und Messfehlern, die aus den Kalibrierungsdaten von ibmq_ehningen bestimmt werden, auch die korrelierten Fehlerraten für Einzel- und Zwei-Qubit Quantengattern aus den Charakterisierungsprotokollen von AP1 verwendet. Hierbei wird die folgende Kostenfunktion verwendet:

$$c = 1 - \prod(1 - s_q) \prod(1 - d_{q,p}) \prod(1 - m_q),$$

wobei m_q den erwarteten Fehler eines Messoperators auf Qubit q beschreibt, s_q den erwarteten Fehler eines Einzel-Qubit-Quantengatters s auf Qubit q beschreibt und $d_{q,p}$ den erwarteten Fehler eines Zwei-Qubit-Quantengatters zwischen dem Qubitpaar d, p beschreibt. Die Fehlerraten der Quantengatter werden weiter bestimmt durch:

$$d_{q,p} = \begin{cases} \mathcal{R}_{w,d_{q,p}}, \text{falls benachbartes Qubit } w \text{ aktiv} \\ \xi_{d_{q,p}} \qquad \text{sonst} \end{cases},$$

$$s_q = \begin{cases} \mathcal{R}_{p,w,s_q} \text{ falls benachbartes Qubit-Paar } p,w \text{ aktiv} \\ \xi_{s_q} \qquad \text{sonst} \end{cases},$$

mit den Kalibrierungsdaten ξ und den Charakterisierungdaten der korrelierten Fehler (AP1) $\mathcal{R}$. D.h. falls die benachbarten Qubits oder Qubit-Paare inaktiv sind, werden die Standard-Kalibrierungsdaten der von ibmq_ehningen verwendeten. Andernfalls treten korrelierte Fehler auf, deren Größenordnung aus den Charakterisierungsdaten (AP1) bestimmt werden. Die Platzierungsmöglichkeit mit den geringsten Kosten c wird für die Berechnung der Quantenschaltung ausgewählt.

Das entwickelte Transpilationsverfahren wurde auf eine Menge von Quantenschaltungen angewandt, die schließlich auf ibmq_ehningen ausgeführt wurden. Die untersuchten Quantenschaltungen bilden arithmetische Funktionen wie das Exklusive-Oder, Quanten-Volume, QAOA und die Hadamard-Leiter ab. Die Ergebnisse der Quantenschaltungs-ausführungen werden in Abbildung 2.10 aufgeführt, wobei die *X*-Achse die verschiedenen Quantenschaltungen beschreibt und die *Y*-Achse die Ergebnisgüte (Hellinger-Güte) der Ausführung mit einem Wert zwischen 1.0 (perfekte Übereinstimmung mit Simulationsergebnissen) und 0 (keine Übereinstimmung mit Simulationsergebnissen) beschreibt. Hierbei wurde in dem entwickelten Transpilierungsverfahren zunächst auf die genaueren Daten durch die Charakterisierung von korrelierten Fehlern verzichtet (Standard, in blau) und schließlich, je nach Aktivitäts-Status der benachbarten Qubits, auf genauere Charakterisierungsdaten aus AP1 zugegriffen (orange).

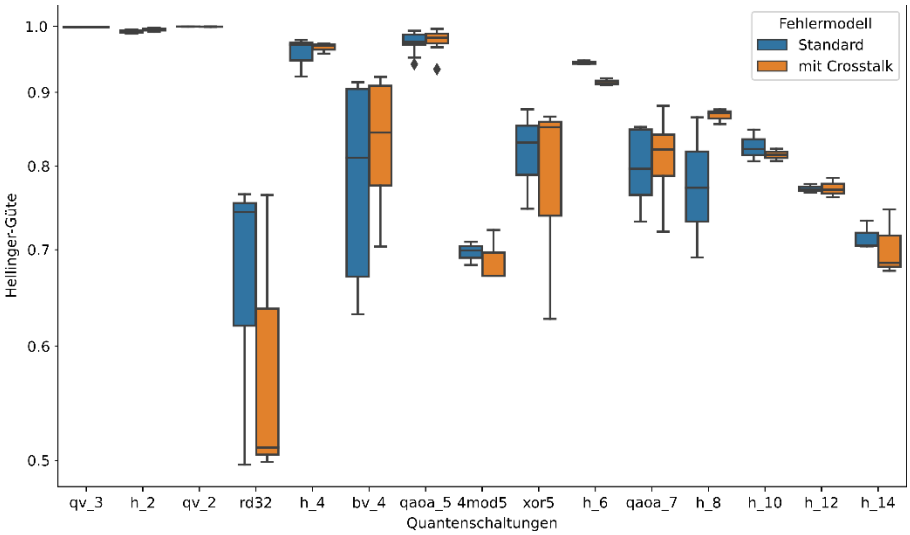


Abbildung 2.10: Ergebnisse der Transpilation mit korrelierten Fehlern

In vielen Fällen ließ sich eine Verbesserung durch das Berücksichtigen von korrelierten Fehlern erreichen (h_8, h_12, xor5). Allerdings gibt es auch einige Quantenschaltungen wie qaoa_5, qv_2, und h_2, für die sich die Ergebnis-Güte kaum unterscheidet oder sich sogar verschlechtert (rd32, h_10). Wir erhoffen uns durch eine genauere Charakterisierung von korrelierten Fehlern, wie im Projekt QORA II geplant, eine a priori Unterscheidung dieser Fälle, worauf ein verbessertes Transpilierungsverfahren Bezug nehmen kann.

2.3 AP 3: Fehlerresistente Algorithmen

Das dritte Arbeitspaket hatte es zum Ziel, die Bestandteile des QAOA-Algorithmus zu analysieren und diese gegebenenfalls durch verbesserte, fehlerresistentere Ansätze zu ergänzen. Einerseits wurden dabei Verbesserungen der zu Grunde liegenden Quantenschaltkreise erreicht, die im fehlerfreien Fall den QAOA-Algorithmus und damit das Portfolioproblem besser lösen als es mit Standardvarianten des Algorithmus möglich ist (siehe Kap. 2.3.1). Auch der klassische Teil des QAOA-Algorithmus, welcher für die Optimierung der Schaltkreisparameter zuständig ist, wurde durch die Untersuchung verschiedener gradientenfreier Optimierungsverfahren verbessert (siehe Kap. 2.3.2). Implementierungen des QAOA-Algorithmus unter der Berücksichtigung von Fehlern zeigten jedoch, dass der im fehlerfreien Fall optimale Algorithmus nicht zwingend zu guten Ergebnissen führt (siehe Kap. 2.3.4). Aus diesem Grund haben wir eine Reihe von Fehlerreduzierungs- und Fehlermitigationsmethoden implementiert und auf den QAOA-Algorithmus angewendet, was zu deutlichen Verbesserungen der Ergebnisse führte (siehe Kap. 2.3.5). Eine Reihe der in AP3 erreichten Ergebnisse wurden in [6] veröffentlicht.

2.3.1 Test und Erweiterung des Quantum Alternating Operator Ansatzes (AP3.1)

Ziel von Arbeitspaket 3 („fehlerresistente Algorithmen“) war die Weiterentwicklung von Quantenoptimierungsalgorithmen, so dass sie resistenter gegenüber Fehlern werden. Im Zentrum steht hierbei der „Quantum Approximate Optimization Algorithm“ bzw. als dessen Verallgemeinerung der „Quantum Alternating Operator Ansatz“. Letzteren bezeichnen wir im Folgenden als „QAOA“ (ersteren als „Standard-QAOA“). Als erstes haben wir in AP3.1 verschiedene Versionen dieses Ansatzes miteinander verglichen. Allen Versionen gemeinsam ist die allgemeine Form des QAOA-Schaltkreises, wonach auf dem Quantencomputer (bzw. dem Simulator) ein Quantenzustand $|\psi\rangle$ von folgender Form präpariert wird:

$$|\psi\rangle = U_M(\beta_p)e^{-i\,\gamma_p F} \dots U_M(\beta_2)e^{-i\,\gamma_2 F}\, U_M(\beta_1)e^{-i\,\gamma_1 F}|\psi_0\rangle$$

wobei F die zu minimierende Funktion bezeichnet, und der sogenannte „Mixer“ U_M je nach Version unterschiedlich gewählt wird. Durch Messungen an diesem Zustand wird der Erwartungswert $\langle F\rangle$ bestimmt. In Verbindung mit einem klassischen Computer werden die Parameter β_i und γ_i $(i = 1,2,\dots,p)$ variiert und hierdurch $\langle F\rangle$ minimiert. In der Standardversion entspricht der Mixer U_M Rotationen einzelner Qubits, wodurch alle Zustände miteinander vermischt werden. Der aus 2-Qubit-Operationen bestehende XY-Mixer mischt hingegen nur solche Zustände, die eine bestimmte Nebenbedingung erfüllen. In unserem Fall besteht diese daraus, dass aus einer gegebenen Liste von Aktien eine bestimmte Anzahl (z.B. 5 aus 10) ausgesucht werden soll. Je nachdem, auf welche Paare von Qubits und in welcher Reihenfolge die Operationen des XY-Mixers angewendet werden, ergeben sich mehrere Untervarianten, von denen die Varianten „xy full“, „xy ring“ und „xy parity ring“ in der Literatur schon beschrieben, aber zu Projektbeginn noch nicht für den Anwendungsfall der Portfoliooptimierung getestet und miteinander verglichen worden sind. Eine weitere Variante „xy full reshuffled“ haben wir in QORA selbst entwickelt. Diese ähnelt der Variante „xy full“, benötigt aber ¼ weniger CNOT-Gatter.

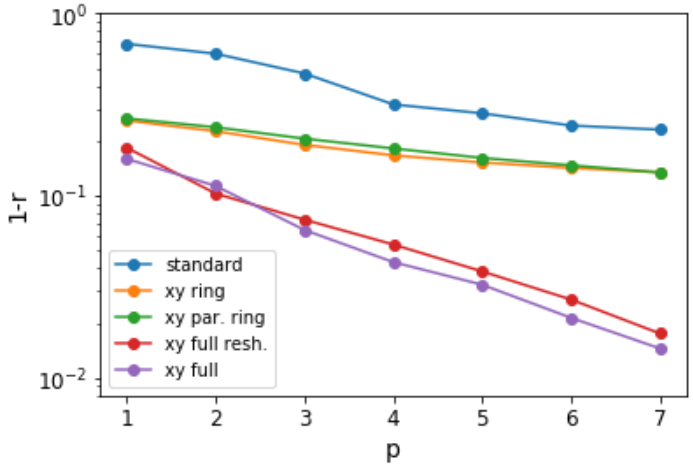


Abbildung 3.1: Vergleich verschiedener QAOA-Varianten für ein Portfoliooptimierungsproblem mit Auswahl von 5 aus 10 Aktien (entsprechend 10 Qubits). Aufgetragen ist die Abweichung $1 - r$ von der optimalen Lösung in Abhängigkeit der Tiefe p des QAOA-Algorithmus für verschiedene Mixer (Standardmixer sowie 4 Versionen des XY-Mixers). Die Rechnungen wurden auf dem „statevector simulator" ausgeführt, welcher den zu minimierenden Erwartungswert fehlerfrei und ohne statistische Schwankungen berechnet. Die Versionen „xy full" (pink) sowie die von uns selbst entwickelte Version „xy full reshuffled" (rot) liefern die besten Ergebnisse, da sie mit steigender Tiefe p am schnellsten gegen die optimale Lösung ($1 - r \to 0$, d.h. approximation ratio $r = 1$) konvergieren.

Die effiziente Implementierung des QAOA-Algorithmus in diesen unterschiedlichen Varianten erfordert es, eine Reihe von Details zu beachten, die in der Literatur nicht in allen Einzelheiten beschrieben sind und hauptsächlich die Optimierung der Parameter β und γ betreffen (z.B. die Wahl geeigneter Anfangswerte, welche der Optimierungsroutine übergeben werden). Um das hierfür nötige technische Know-How aufzubauen, waren unsere Vorarbeiten aus dem Startprojekt Quantencomputing hilfreich. Die Ergebnisse unseres Vergleichs verschiedener Mixer (Meilenstein M6) für die Lösung eines Portfoliooptimierungsproblems mit $n = 10$ Aktien (entsprechend 10 Qubits) auf einem fehlerfreien Simulator („statevector simulator") sind in Abbildung 3.1 zu sehen. Die Varianten „xy full" und „xy full reshuffled" schneiden hierbei am besten ab. Ähnliche Ergebnisse erzielen wir auch mit dem „qasm simulator", welche die durch eine endliche Anzahl von Messungen („shots") verursachten statistischen Schwankungen beinhaltet. Hierfür ist es jedoch eine andere Optimierungsroutine nötig, welche mit diesen Schwankungen umgehen kann.

2.3.2 Gradientenfreie klassische Optimierungsroutinen für QAOA (AP3.2)

Ein wichtiger Schritt im QAOA Algorithmus ist die klassische Optimierung der Parameter γ,β, die respektive Evolutionszeiten für den Zielhamiltonian H_C und den Mischerhamiltonian H_B sind. Die Optimierung hat die Minimierung des Erwartungswertes der Energie in einem Endzustand zum Ziel, der abhängig von diesen Parametern vom QAOA erzeugt wird. Dieser Erwartungswert kann aber i.A. nur aus Messergebnissen geschätzt werden, wobei sehr viele Realisierung des Algorithmus nötig sind, die auf zufällig verteilte Messergebnisse führen. Dementsprechend ist die Zielfunktion i.A. erheblich verrauscht, und direkte Gradientenverfahren zum Auffinden der optimalen Parameter γ,β geraten in solch verrauschten Energielandschaften in Schwierigkeiten.

Ziel des AP3.2 war daher, alternative klassische Optimierungsverfahren für ihre Eignung in QAOA zu untersuchen. Dies kann aber nicht unabhängig vom Rest des Algorithmus geschehen, und auch nicht unabhängig von der Art des untersuchten Problems. Insbesondere der gewählte Mischerhamiltonian hat großen Einfluss auf die

Leistungsfähigkeit des Algorithmus, ebenso wie mögliche Initialisierungen der Parameter γ,β. Und auch von der Schwierigkeit des Problems hängt die beste mögliche Leistungsfähigkeit ab. Daher haben wir eine umfangreiche Untersuchung der Leistungsfähigkeit für eine große Zahl von Kombinationen von Mischerhamiltonian, klassischen Optimierungsverfahren, Parameterinitialisierungen und Schwierigkeiten des Problems untersucht. In den Simulationen besteht zudem die Wahl zwischen Zustandsvektorsimulationen oder QASM Simulationen. Letztere sind besser an die Hardware angepasst und erlauben es, einzelne Durchläufe zu simulieren, wo jeder Durchlauf zu einem der zufällig verteilten Messergebnisse führt.

Zunächst untersuchten wir mittels Zustandsvektorsimulation die klassischen Optimierungsverfahren GD („gradient descent"), SGD („stochastic gradient descent"), NMEAD („Nelder Mead"), COBYLA („Constrained optimization by linear approximation"), und SLSQP („Sequential least squares programming algorithm"). Diese wurden kombiniert mit dem standard mixer, xy_full_mixer, xy_reshuffled_mixer, xy_ring_mixer and xy_parity ring mixer (siehe Kap. 2.3.1). SLSQP zeigte vielversprechende Resultate mit allen Mischern für einfache Probleme mit 5 oder 10 Aktien. Sowohl SLSQP als auch NMEAD und COBYLA erzielten um die 99% approximation ratio (AR) und 90% Grundzustandswahrscheinlichkeit mit dem „xy full mixer" (GZW). Dagegen funktionieren die Gradientenverfahren GD und SGD nur mit dem Standardmischer einigermaßen vernünftig.

Der Blick auf diese Resultate wird aber modifiziert durch die Beobachtung, dass die Leistung des SLSQP auf dem QASM Simulator deutlich reduziert wird. Unabhängig vom Mischer wurden dort nur noch ca. 55-60% AR erzielt, während COBYLA ca. 95% AR und 70% GZW in Kombination mit dem xy_full_mixer bei 1000 Realisierungen eines 5-Aktien Problems. Sprich, die gradientenfreien Verfahren zeigten in den realistischeren QASM Simulationen tatsächlich bessere Leistung.

Zusätzlich zu diesen im Projekt geplanten Untersuchungen studierten wir die Variante „warm start" QAQA, und warm start QAOA mit dem Standard-Mischer. Beim warm start fängt man mit einem Anfangszustand an, der aus der Relaxation des Problems mit Binärzahlen auf reelle Zahlen gewonnen wird, welche dann wieder auf die nächsten Binärzahlen reduziert werden, was durch die Tatsache motiviert wird, dass das relaxierte Problem effizient klassisch gelöst werden kann (quadratische Optimierung). Bei 10 assets erzielte WSQAOA bereits AR=90% mit COBYLA und gerade mal p=2 Iterationen, während hier der Standard QAOA nur auf ca. AR=50% kommt.

2.3.3 Leichte und schwere Portfoliooptimierungsprobleme

Parallel zu diesen Studien wurde u.a. die Abhängigkeit der Leistung des QAOA von der Verteilung der Korrelationen sowie Gewinne der Aktien eines Korbes untersucht. Motiviert wurde dies durch die Beobachtung, dass AR sowie GZW über mehrere Optimierungsläufe jeweils zufällig erzeugter Körbe stark variierten. Um Instanzen mit besonderer Ausprägung dieses Effekts zu finden, optimierten wir 2400 zufällig generierte Körbe. Daraufhin untersuchten wir Instanzen, bei welchen der QAOA eine besonders gute/schlechte AR sowie GZW aufwies, auf instanzübergreifende Merkmale. Hier fiel u.a. auf, dass Korrelationen sowie Gewinne der Aktien aus Körben guter und schlechter Optimierungsleistung unterschiedlich stark fluktuierten. Dieses Muster trat sowohl bei Körben mit Aktien des Aktienindex DAX30, als auch MDAX50 auf. Eine mögliche Erklärung hierfür liegt in den Energielandschaften der jeweiligen Instanzen. Eine starke Fluktuation der Korrelationen sowie Gewinne wird zu einer starken Fluktuation der Energielandschaft propagiert. Dies erleichtert die klassische Optimierung und führt so zu einem besseren Ergebnis. Analog lässt sich eine schlechte Leistung des QAOA für kleine Fluktuationen erklären.

2.3.4 Evaluation des QAOA-Algorithmus unter Fehlern (AP3.3)

Für die Untersuchungen der Konvergenz wurden verschiedene Varianten von QAOA unter Fehlern simuliert, bei denen ein Teil der Ansatzschaltung, der sogenannte *Mixer*, angepasst wurde (solche Lösungen sind von anderen Projektpartnern vorgeschlagen worden). Dabei sind teilweise unerwartete Ergebnisse beobachtet worden: die *XY-Mixer*,

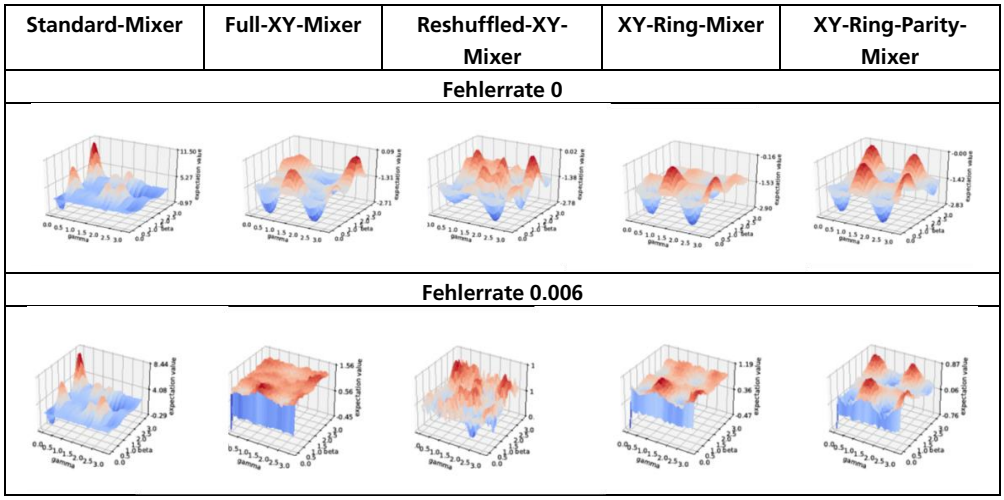


Abbildung 3.2: Simulierte Optimierungslandschaften im fehlerfreien Fall und unter Annahme einer moderaten Fehlerrate (Depolarisationsfehler).

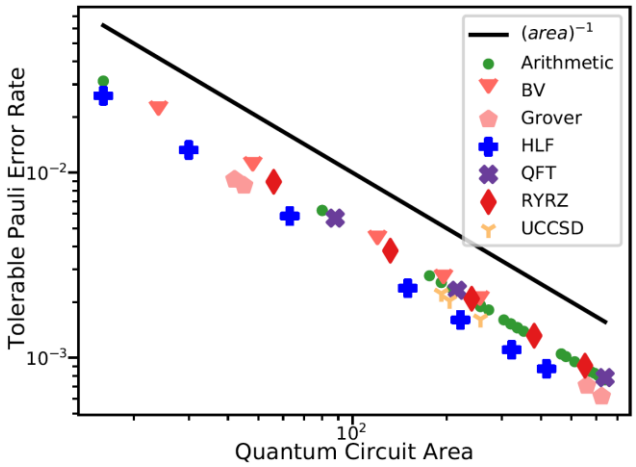


Abbildung 3.3: Zusammenhang zwischen der Größe einer Quantenschaltung und der maximalen Fehlerrate, mit der diese Schaltung betrieben werden kann. Bild aus der Veröffentlichung [1].

die ohne Fehler dem Standard-QAOA überlegen sind, werden deutlich stärker von Fehlern beeinträchtigt [6]. Die Optimierungslandschaften in Abb. 3.2 illustrieren dieses Ergebnis: Unter Fehlern sind die Landschaften deutlich „zerklüfteter", der Algorithmus hat also größere Schwierigkeiten, zu konvergieren und das Optimum zu finden. Diese Ergebnisse stellten eine Grundlage für weitere Arbeiten in AP3 sowie AP4 dar.

Die Untersuchungen zur Schaltungsgröße haben zu einer gemeinsamen Publikation [1] geführt und werden in Abb. 3.3 zusammengefasst. Im Wesentlichen ist bei einer festgelegten Erfolgswahrscheinlichkeit eine reziproke Beziehung von Schaltungsgröße und unterstützter Fehlerrate beobachtet worden. Allerdings sind darüber hinaus Effekte

beobachtet worden, die auf Unterschiede in der Struktur der Schaltung zurückzuführen sind und die Erfolgswahrscheinlichkeit beeinflussen. Diese Ergebnisse sollen bei der Erhöhung der Fehlerresistenz von QAOA in AP3 berücksichtigt werden.

2.3.5 Implementierung der Fehlerreduzierung (AP3.4) und verbesserter QAOA-Algorithmus unter Fehlern (AP3.5)

Um die auf dem IBMQ System real auftretenden Fehler so weit wie möglich zu reduzieren, wurden diverse Methoden der Fehlerreduzierung und Fehlermitigation implementiert. Hierzu wurden schon in der ersten Projekthälfte einige Schritte unternommen. Zum Beispiel wurde eine in AP2.2 optimierte Transpilation des QAOA-Schaltkreises verwendet, welche mit möglichst wenig CNOT-Gattern auskommt. Bei der Auswahl der Qubits (4 von 27) wurden die Fehlerraten der beteiligten CNOT-Gatter berücksichtigt. Weiterhin haben wir eine in Qiskit bereitgestellte Methode zur Mitigation von Auslesefehlern benutzt. Mit Hilfe dieser Methoden wurde die Optimierung des QAOA-Algorithmus durch Simulationen und Ausführungen auf dem IBMQ System untersucht. Dies führte zu den Resultaten welche in Kap. 2.4.2 und entsprechend Abbildung 4.1 vorgestellt werden.

Im zweiten Teil des Projekts haben wir die Methoden zur Fehlerreduzierung weiter verbessert. Zunächst haben wir mittels verbesserter Transpilation (siehe Kap. 2.2.1) die Anzahl der für die QAOA-Schaltkreise benötigten CNOT-Gatter im Vergleich zu oben um 2 reduziert. Weiterhin wurden bei der Auswahl der Qubits nun auch die Daten aus der Crosstalk-Charakterisierung (Kap. 2.1.2) verwendet, um die Fehlerraten inklusive Crosstalk möglichst gering zu halten. Des Weiteren haben wir auch die in Qiskit bereitgestellte Version von Dynamical Decoupling aktiviert: Bleibt ein Qubit längere Zeit ungenutzt (weil erst Operationen auf anderen Qubits abgewartet werden müssen), werden währenddessen auf dieses Qubit zu geeignet gewählten Zeitpunkten zwei X-Pulse angewendet, um eventuell auftretende Dephasierungseffekte zu reduzieren. Die Resultate des QAOA-Algorithmus wurden unter Ausnutzung dieser Methoden generell verbessert. Entsprechend wurde in dem simulierten Fehlermodell Crosstalk berücksichtigt (Kap. 2.1.3), was zu einer besseren Übereinstimmung mit den auf ibmq_ehningen erhaltenen Ergebnissen führt (blaue und grüne Balken in Abb. 4.2 im Vergleich zu Abb. 4.1 in Kap. 2.4.2). Auch im Falle des IBMQ Systems führen die auf diese Weise erhaltenen Ergebnisse zu einer deutlichen Verbesserung (siehe Abb. 4.2 in Kap. 2.4.2), d.h., zu einer höheren Wahrscheinlichkeit des optimalen Zustands. Insgesamt können wir somit einen fehlerresistenteren QAOA-Algorithmus vorweisen (Meilenstein M8).

2.4 AP4: Portfoliooptimierung

Das Arbeitspaket 4 hat die Zielsetzung, die wissenschaftlichen Erkenntnisse und Fortschritte der AP1 bis 3 anhand eines realitätsnahen Use-Cases zu demonstrieren. Hierfür wurde das Problem der Portfoliooptimierung nach Markowitz (1952) als eines der klassischen NP-vollständigen Optimierungsprobleme der Finanzwissenschaft ausgewählt. Wie nachfolgend dargestellt, konnte anhand dieses Beispiels eine für das Quantumcomputing geeignete Formulierung gefunden werden (siehe Kap. 2.4.1). Des Weiteren konnten für kleine Probleminstanzen die Lösungen des Quantensimulators mit der echten Quantenhardware verglichen werden. Insbesondere konnten wir durch Ausnutzung der in AP1-3 entwickelten Methoden die Qualität dieser Lösungen im Laufe des Projekts deutlich steigern (siehe Kap. 2.4.2). Schließlich diente das Anwendungsbeispiel der Portfoliooptimierung auch als Ausgangspunkt für die Abschätzung des Quantenvorteils (siehe Kap. 2.4.3).

2.4.1 Definition von Prototypen für die Portfoliooptimierung (AP4.1)

Als erste Maßnahme in AP4 wurde der zu betrachtende Use Case Portfoliooptimierung ausgearbeitet und mit Hilfe von realen Marktdaten formuliert. Die ursprünglich auf zurückgehende Formulierung des Problems der Portfoliooptimierung diente dabei als Ausgangspunkt. Die klassische Formulierung – ursprünglich für kontinuierliche Portfoliogewichte vorgeschlagen – wird dabei so angepasst, dass nunmehr binäre Portfoliogewichte gesucht werden. D.h. eine Aktie befindet sich im Portfolio (Zustand 1) oder sie befindet sich nicht im Portfolio (Zustand 0). Dabei gilt die Nebenbedingung, dass eine fest vorgegeben Anzahl von Aktien für das Portfolio ausgewählt werden soll. Mit binären Portfoliogewichten kann dann das Ursprungsproblem als sog. QUBO (Quadratic Unconstrained Binary Optimization) Problem umgeschrieben werden. Ein QUBO-Problem wiederum ist durch die Ersetzung der binären Portfoliogewichte durch Qubits eine für einen Quantencomputer zugängliche Problemklasse. Durch eine Zerlegung von Integer-Zahlen als Vielfache von 2-er Potenzen konnte des Weiteren eine Formulierung der Portfoliooptimierung gefunden werden, die nicht nur binäre Werte erlaubt, sondern auch ganzzahlige Werte. Zwar werden in dieser Formulierung für jede Aktie mehrere Qubits verwendet, die Formulierung ist aber als erheblich praxisgerechter anzusehen. Es konnte gezeigt werden, dass sich bei der Verwendung von ganzzahligen Portfoliowerten die strukturellen Eigenschaften des Problems nicht ändern. Eine andere Verbesserung des ursprünglichen Ansatzes der Portfoliooptimierung betrifft die Berücksichtigung von Transaktionskosten. Hierbei wird die Annahme getroffen, dass ausgehend von einem gegebenen Portfolio jede Um-Allokation mit Kosten (z.B. Gebühren für eine Börse oder einen Broker) verbunden ist. Auch unter dieser Erweiterung bleibt die QUBO-Struktur invariant. Es ändern sich lediglich die die Parameter und dadurch ggf. die Performanz der Lösungsalgorithmen auf dem Quantencomputer.

Für die Portfoliooptimierung werden zum einen die Kovarianzmatrix und zum anderen der Renditevektor der betrachteten Aktienwerte benötigt. Diese Daten wurden anhand realer Marktdaten aus diversen Marktsegmenten (DAX, MDAX, EuroStoXX sowie ETFs) ermittelt. Durch die Verwendung unterschiedlicher Marktsegmente lässt sich untersuchen, inwieweit die Lösungsgüte der Quantenalgorithmen von den Parametern des Problems (Kovarianzmatrix und Renditevektor) abhängt. Da reale Markdaten in der Regel verrauscht sind bzw. Fehlstellen besitzen, wurden im Laufe des Projekts unterschiedliche De-Noising-Methoden untersucht, um reale Marktdaten zu bereinigen. Diese Methoden umfassten zum einen etablierte Verfahren aus der Random Matrix Theorie, aber auch Datenbereinigungsverfahren auf Basis tiefer neuronaler Netze.

Weitere Anwendungsfälle

Im Laufe des Projekts wurde – auch durch Diskussionen mit dem Industriebeirat – untersucht, inwieweit andere für den Finanzbereich relevante Fragestellungen mit dem dargestellten QUBO-Formalismus gelöst werden können. Als erste Problemstellung wurde die Merkmalsauswahl bei einem Klassifikationsalgorithmus wie er z.B. für das Kreditscoring verwendet wird, aufgegriffen und untersucht. In der Praxis ist es regelmäßig wichtig, dass ein Klassifikationsalgorithmus auf der einen Seite eine hohe Trennschärfe besitzt, auf der anderen Seite hierfür aber eine möglichst geringe Anzahl von Merkmalen benötigt. Obwohl sich diese fachliche Problemstellung von der Portfoliooptimierung unterscheidet, führt sie auf eine QUBO-Formulierung, die sich mit den Methoden und Verfahren unseres Projekts lösen lässt. In Analogie zur Untersuchung unterschiedlicher Marktsegmente bei der Portfoliooptimierung ist auch eine relevante Fragestellung, inwieweit die Lösungsgüte der Quantenalgorithmen von den genauen Parametern des Problems abhängen.

2.4.2 Implementierung auf dem Quantencomputer (AP4.2, AP4.3)

Für die Lösung des als QUBO formulierten Portfoliooptimierungsproblems wurden unterschiedliche Quantenalgorithmen angewendet. Zum einen der QAOA-Ansatz (der im Fokus des QORA-Projektes steht), zum anderen der VQE-Ansatz als mögliche Alternative. Für die Implementierung der Algorithmen wurden verschiedene, von IBM entwickelte QISKIT-Module verwendet. Diese Module erlauben auf der einen Seite eine Problemlösung „out of the box" und auf der anderen Seite eine eigenständige Implementierung auf der Ebene von Quantengattern. Die unterschiedlichen Berechnungen wurden mit Quantensimulatoren durchgeführt und dienten dazu, Benchmarkwerte in einem idealen Umfeld zu ermitteln. Als Performanzmaß diente dabei jeweils die sog. Approximation Ratio, d.h. wie gut die Quantenalgorithmen in der Lage sind, die auf klassischem Wege ermittelte Lösung zu reproduzieren. Eine wichtige Frage an dieser Stelle war außerdem, inwieweit in der Literatur vorgeschlagene Verbesserungen von QAOA bzw. VQE (z.B. CVAR-Optimization oder Warm-Start-Approach) für das hier betrachtete Problem eine Verbesserung bringen.

Portfoliooptimierung auf ibmq_ehningen (AP4.2)

Auf dem echten Quantencomputer können aufgrund der dort auftretenden Gatter- und Auslesefehler (siehe Kap. 2.1) bislang nur relativ kleine Probleminstanzen behandelt werden. In Abbildung 4.1 zeigen wir daher unsere Ergebnisse für ein im Vergleich zu Abbildung 3.1 reduziertes Portfoliooptimierungsproblem mit 4 anstelle von 10 Qubits und QAOA-Tiefe $p = 3$ statt 7. Um die Fehler möglichst weitgehend zu reduzieren, haben wir die in Kap. 2.3.4 diskutierten Methoden zur Fehlerreduzierung und -mitigation verwendet. Abbildung 4.1 zeigt die erzielten Ergebnisse (Meilenstein M9), welche auf Basis der Arbeiten in der ersten Projekthälfte erzielt wurden, als grüne Balken. Der Vergleichswert 1/16 für die Wahrscheinlichkeit des optimalen Portfolios bei rein zufälliger Auswahl wird in allen Fällen überschritten, was zeigt, dass die Optimierung im Prinzip funktioniert. Dennoch ist vor allem im Fall der theoretisch überlegenen „xy full"-Variante (Mitte) der Einfluss der Fehler im Vergleich zu den roten Balken klar erkennbar.

Demonstration der verbesserten Portfoliooptimierung (AP4.3)

In der zweiten Projekthälfte ist es uns gelungen, die oben in Abbildung 4.1 gezeigten Ergebnisse, deutlich zu verbessern, siehe Abbildung 4.2. Im Falle des IBMQ Systems (grüne Balken) konnten wir durch die Implementierung weiterer Fehlerreduzierungsmethoden, wie etwa Dynamical Decoupling (siehe Kap. 2.3.4), eine höhere Wahrscheinlichkeit für die optimale Lösung erhalten. Auch die Simulationen des IBMQ Systems (blaue Balken), welches auf dem neuen in Kap.2.1 eingeführten Crosstalk-

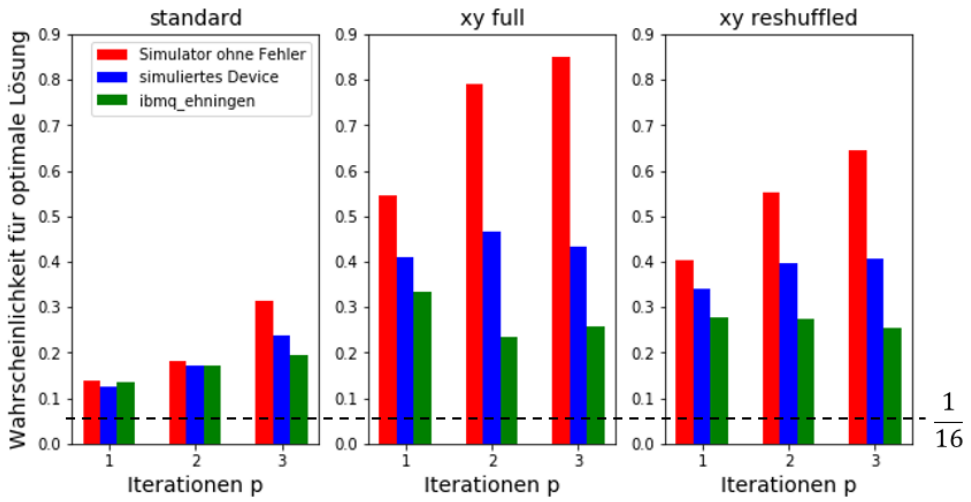


Abbildung 4.1: Vergleich der Performance von Simulatoren und IBM-Quantencomputer (ibmq_ehningen) bei der Berechnung eines optimalen Portfolios von 2 aus 4 Aktien (entsprechend 4 Qubits). Gezeigt ist für drei verschiedene QAOA-Varianten (links: „standard", Mitte: „xy full", rechts: „xy reshuffled", siehe Kap. 0) die Wahrscheinlichkeit, nach QAOA-Optimierung der Tiefe p die optimale Lösung zu messen, und zwar auf dem fehlerfreien Simulator (rote Balken), auf dem Simulator mit dem in Qiskit verfügbaren Fehlermodell von ibmq_ehningen (blaue Balken) und schließlich auf dem „echten" Device (grüne Balken). Der schon in Abbildung 5 gezeigte theoretische Vorteil der XY-Mixer im fehlerfreien Fall (rote Balken) ist auf dem fehlerbehafteten Device (grüne Balken) weniger stark ausgeprägt. Dennoch ist auch dort die Optimierung nicht wirkungslos: Die Wahrscheinlichkeit, das optimale Portfolio zu erhalten, überschreitet in allen Fällen deutlich den Vergleichswert 1/16 für ein aus 4 Aktien zufällig ausgesuchtes Portfolio. Die Abweichungen zwischen dem simulierten Fehlermodell (blaue Balken) und dem wirklichen Quantencomputer ibmq_ehningen (grüne Balken) sind vermutlich auf im Modell nicht berücksichtigten Crosstalk zwischen gleichzeitig durchgeführten Gattern (siehe Kap. 2.1.2) zurückzuführen.

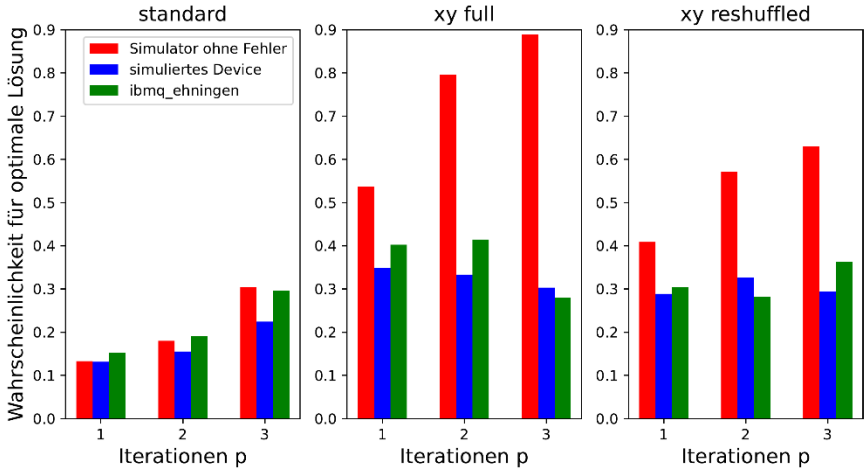


Abbildung 4.2: Aktualisierte Version von Abbildung 4.1, welche im Vergleich damit die Fortschritte im zweiten Projektjahr veranschaulicht. Folgende Verbesserungen wurden hierfür implementiert: Reduktion der Anzahl der CNOT-Gatter mit einem verbesserten Transpilationsverfahren (siehe Kap. 2.2.1), Berücksichtigung von Crosstalk-Daten bei der Auswahl der Qubits von ibmq_ehningen (siehe Kap. 2.1.3) und Verwendung von Dynamical-Decoupling-Pulssequenzen. Die auf ibmq_ehningen erhaltenen Wahrscheinlichkeiten der optimalen Lösung (grüne Balken) sind durchgängig größer als in Abbildung 4.1, außerdem stimmen sie besser mit den Ergebnissen der durch Crosstalk-Effekte ergänzten Simulation (blaue Balken) überein.

empfindlichen Fehlermodell beruhen, zeigen eine bessere Übereinstimmung mit den auf ibmq_ehningen erhaltenen Ergebnissen. Diese Verbesserungen spiegeln die bereits in Kapitel 2.1 und 2.2 diskutierten Fortschritte in der Fehlercharakterisierung, -modellierung und -reduzierung wieder.

2.4.3 Abschätzung des Quantenvorteils (AP 4.4)

Die Abschätzung eines möglichen Quantenvorteils von QAOA gegenüber klassischen Optimierungsverfahren stellte sich schwieriger dar als ursprünglich im Antrag vorhergesehen. Während dort die Resilienz von QAOA gegenüber Fehlern thematisiert wird, ist nämlich bislang nicht einmal für den fehlerfreien Fall wirklich nachgewiesen, dass ein derartiger Quantenvorteil wirklich existiert. In der Literatur findet man nur wenige Arbeiten, in denen QAOA direkt mit klassischen Verfahren zur Lösung des kombinatorischen Optimierungsproblems verglichen wird: Eine davon [Guerreschi/Matsuura, Sci. Rep. **9**, 6903 (2019)] liefert Indizien für einen Quantenvorteil im Bereich mehrerer hundert Qubits, allerdings ist die Extrapolation auf viele Qubits dort nicht sorgfältig durchgeführt (insb. weil die Iterationstiefe p dabei konstant gehalten wird). Ein weiterer Artikel [Lykov *et al.*, arXiv:2206.03579 (2022)] betont die Leistungsfähigkeit klassischer Algorithmen und setzt dabei enge Grenzen für das Erreichen eines Quantenvorteils. Beide Arbeiten betrachten das MaxCut-Problem mit 3 Verbindungen pro Knoten. Vergleichbare Abschätzungen für andere Optimierungsprobleme, insbesondere Portfoliooptimierung, sind uns nicht bekannt. Dies ist umso erstaunlicher, als dass die Lösung von kombinatorischen Optimierungsproblemen häufig als mögliches Anwendungsfeld von Quantencomputern genannt wird.

Mit dem Ziel, diese wichtige Lücke zu schließen, haben wir die Portfoliooptimierung mittels QAOA zunächst für den fehlerfreien Fall mit einem klassischen Optimierungsverfahren verglichen. Hierzu betrachten wir jeweils 10 zufällig gewählte Instanzen (basierend auf realen Marktdaten aus dem DAX, siehe Kap. 2.4.1) von Portfoliooptimierungsproblemen verschiedener Größe ($N = 10$ bis 28 Aktien, von denen $\lfloor N/2 \rfloor$ auszuwählen sind) und bestimmen jeweils mittels „Brute Force"-Suche die optimale Lösung. Dann verwenden wir einen klassischen Algorithmus zur Lösung von QUBO-Problemen (die Methode „Burer" aus der Bibliothek MQLib, welche im oben erwähnten Artikel von Lykov die besten Ergebnisse erzielt) und bestimmen die Zeit, innerhalb der dieser Algorithmus die optimale Lösung findet. Wie in Abbildung 4.3 (linke Seite) zu sehen, treten zwischen den 10 für jeden Wert von n betrachteten Instanzen erhebliche Schwankungen auf (blaue Punkte), wobei im (geometrischen) Mittel (blaue Linie) die Zeit zum Auffinden der optimalen Lösung wie $2^{n/3}$ skaliert (rote gestrichelte Linie).

Um nun den Vergleich mit dem QAOA-Algorithmus durchzuführen, schätzen wir die Zeit ab, welche ein Quantencomputer für das Auffinden der optimalen Lösung benötigen würde. Da im Verlauf der Optimierung viele Quantenschaltkreise durchlaufen werden, addieren wir hierzu die Tiefen aller durchgeführten Quantenschaltkreise (jeweils multipliziert mit der Anzahl der Shots), bis die optimale Lösung zum ersten Mal gemessen wird.[1]

[1] Um die durch den Messprozess bedingten statistischen Schwankungen zu reduzieren, ermitteln wir genauer gesagt den Median dieser Tiefe. Dieser ist durch die Bedingung definiert, dass die Wahrscheinlichkeit, die optimale Lösung im Verlauf der Optimierung mindestens einmal gemessen zu haben, genau ½ beträgt.

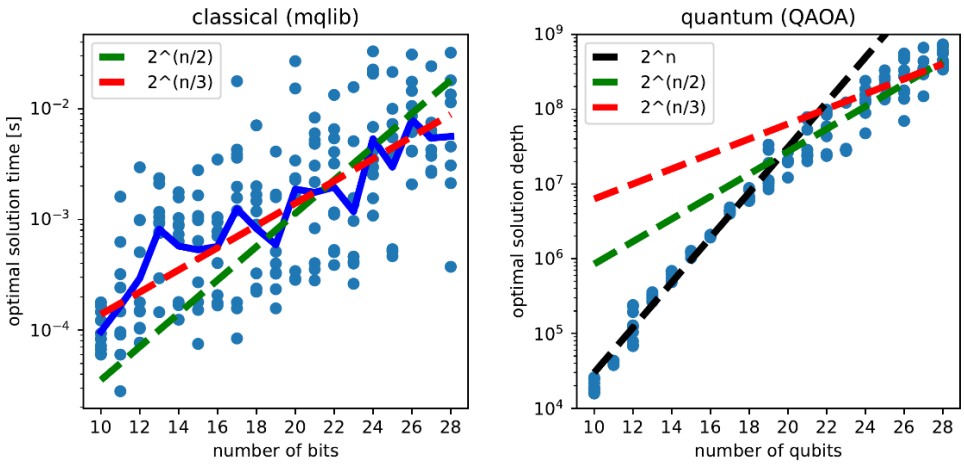


Abbildung 4.3: Vergleich eines klassischen Verfahrens zur Lösung von QUBO-Problemen mit QAOA. Links: vom klassischen Algorithmus benötigte Zeit zum Auffinden der optimalen Lösung als Funktion der Problemgröße n (Anzahl binärer Variablen) für jeweils 10 Instanzen des Portfoliooptimierungsproblems (blaue Punkte, mit blauer Linie als geometrischem Mittel). Die benötigte Zeit skaliert wie $2^{n/3}$ (rote gestrichelte Linie). Rechts: Aufaddierte Gesamttiefe aller Quantenschaltkreise, welche auf dem fehlerfreien Simulator im Verlauf der QAOA-Optimierung (mit Standard-Mixer und 1000 Shots pro Schaltkreis) insgesamt ausgeführt werden, bis die optimale Lösung zum ersten Mal gemessen wird, als Funktion der Problemgröße n (Anzahl der Qubits). Für kleine Probleme beobachten wir eine Skalierung wie 2^n, entsprechend einer „Brute Force"-Suche. Ab ca. $n = 20$ führt die Quantenoptimierung zu einer gemäß $2^{n/2}$ (grüne gestrichelte Linie) beschleunigten Lösung. Die noch bessere Skalierung des klassischen Algorithmus (rote gestrichelte Linie) wird jedoch nicht erreicht. Insgesamt liefert die hier verwendete QAOA-Variante also keinen Hinweis auf einen Quantenvorteil.

Dem liegt die Annahme zu Grunde, dass die Zeit zur Ausführung eines Quantenschaltkreises proportional zu dessen Tiefe ist.[2]

In Abbildung 4.3 (rechts) ist diese Größe für dieselben Probleminstanzen wie auf der linken Seite aufgetragen, um einen direkten Vergleich mit dem klassischen Algorithmus zu erhalten. Für kleine Probleme ($n < 20$) skaliert die Gesamttiefe aller bis zum Auffinden der optimalen Lösung durchgeführten Schaltkreise wie 2^n, was einer zufälligen Suche entspricht. Dies ist dadurch zu erklären, dass der QAOA-Algorithmus einige Zeit benötigt, um optimierte Werte für die Schaltkreisparameter $\vec{\beta}$ und $\vec{\gamma}$ zu finden, wobei jeder Schaltkreis mit 1000 Shots durchgeführt wird. Ab ca. $n = 20$ macht sich die Optimierung der Parameter bemerkbar, so dass die optimale Lösung nun deutlich schneller als bei zufälliger Suche gefunden wird (Skalierung wie $2^{n/2}$). Das Skalierungsverhalten ist jedoch immer noch ungünstiger als beim klassischen Algorithmus, bei dem die zum Auffinden der optimalen Lösung benötigte Zeit weniger schnell mit n steigt ($2^{n/3}$). Insgesamt sehen wir in diesem Fall also keinen Hinweis auf einen Quantenvorteil.

Dieses Ergebnis bestärkt uns in dem Vorhaben, im Nachfolgeprojekt QORA II auch über den Standard-QAOA hinausgehende algorithmische Konzepte der Quantenoptimierung zu verfolgen. Gleichzeitig gilt es auf der Suche nach einem Quantenvorteil solche

[2] Daneben könnte man beispielsweise auch noch eine von der Tiefe des Schaltkreises unabhängige Zeit zur Präparation des Anfangszustands berücksichtigen. In unserer Analyse beschränken wir uns auf die Tiefe, um eine von spezifischen Eigenschaften zukünftiger Quantenhardware unabhängige Abschätzung der Skalierung mit steigender Problemgröße zu erhalten.

Probleme zu identifizieren, die mit klassischen Verfahren schwerer zu lösen sind, als es beim Portfoliooptimierungsproblem der Fall ist.

3 Anhang

3.1 Publikationen

Anzahl der projektbezogenen wissenschaftlicher Publikationen: 9

[1] Sebastian Brandhofer, Simon J. Devitt, Thomas Wellens, and Ilia Polian: *Special Session: Noisy Intermediate-Scale Quantum (NISQ) Computers – How They Work, How They Fail, How to Test Them?* Proc. 39th IEEE VLSI Test Symposium (VTS'21), pp. 1-6 (2021).

[2] Yanjun Ji, Sebastian Brandhofer, and Ilia Polian, *Calibration-Aware Transpilation for Variational Quantum Optimization,* 2022 IEEE Int. Conf. on Quantum Computing and Engineering (QCE), pp. 204-214 (2022).

[3] Jelena Mackeprang, Daniel Bhatti, Matty J. Hoban, and Stefanie Barz, *The power of qutrits for non-adaptive measurement-based quantum computing*, arXiv:2203.12411 (2022).

[4] Joris Kattemölle and Guido Burkard, *Effects of correlated errors on the Quantum Approximate Optimization Algorithm*, arXiv:2207.10622 (2022), als "Editor's Suggestion" zur Veröffentlichung in Physical Review A akzeptiert unter dem Titel "Ability of error correlations to improve the performance of variational quantum algorithms"

[5] Nadia Milazzo, Olivier Giraud, Giovanni Gramegna, and Daniel Braun*, Principles of quantum functional testing*, arXiv:2209.11712 (2022).

[6] Sebastian Brandhofer, Daniel Braun, Gerhard Hellstern, Matthias Hüls, Yanjun Ji, Ilia Polian, Amandeep Singh Bhatia and Thomas Wellens: *Benchmarking the performance of portfolio optimization with QAOA*, Quantum Inf. Process. **22**, 25 (2023).

[7] Yanjun Ji, Ilia Polian, Algorithm-Aware Qubit Mapping, TechRxiv.21786146 (2022)

[8] Yanjun Ji, Kathrin F. Koenig, Ilia Polian, Optimizing QAOA on Bipotent Architectures, arXiv:2303.13109 (2023)

[9] Andreas Ketterer, Thomas Wellens, Characterizing crosstalk of superconducting transmon processors, arXiv:2303.14103 (2023)

3.2 Vorträge

- S. Brandhofer, *Analysing NISQ Compatibility of Quantum Algorithms*, 28. April 2021, IEEE VLSI Test Symposium
- T. Wellens, *Error Characterization on Quantum Hardware*, 28. April 2021, IEEE VLSI Test Symposium
- T. Wellens, *Quantum optimization using resilient algorithms*, 17. Juni 2021, Quantum Business Network Meeting on Quantum Computing for Finance
- T. Wellens, *Quantenoptimierung mit resilienten Algorithmen*, 8. Juli 2021, Fraunhofer Solution Days
- T. Wellens, *Quantencomputing – Einführung und Anwendungspotentiale*, 25. September 2021, Bodensee-Seminar der Industrie- und Handelskammer Hochrhein-Bodensee (Konstanz)
- V. Dehn, *Portfoliooptimierung mit dem Quantencomputer*, 1. Dezember 2021, Quantum Village Ehningen

- J. Kattemölle, *Noise resilience of the Quantum Approximate Optimization Algorithm (QAOA) against (un)correlated errors*, 16. März 2022, APS March meeting, Chicago
- J. Kattemölle, *Effects of correlated errors on the Quantum Approximate Optimization Algorithm*, 8 September 2022, DPG Frühjahrstagung SKM, Regensburg
- A Ketterer, *Gate-Error Characterization via Long-Sequence Quantum Process Tomography*, 8 September 2022, DPG Frühjahrstagung SKM, Regensburg
- J. Kattemölle, *Error correlations can improve the performance of VQAs*, 20. Oktober 2022, KQCBW Developer Conference, Ehningen
- A. Ketterer, *Gate-Error Characterization via Long-Sequence Quantum Process Tomography*, 20. Oktober 2022, KQCBW Developer Conference, Ehningen
- T. Wellens, R. Seidel, *Quantum computing at Fraunhofer: From error characterization to high-level programming*, Practitioners Forum zum IBM Quantum Summit, 10. November 2022, New York
- J. Kattemölle, *Error correlations can improve the performance of variational quantum algorithms*, 10 Januar 2023, 778. WE-Heraeus-Seminar, Physikzentrum Bad Honnef (Poster)
- A. Ketterer, *Quantum gate-error and crosstalk characterization on superconducting transmon processors*, 10. Januar 2023, 778. WE-Heraeus-Seminar, Physikzentrum Bad Honnef (Poster)
- A. Ketterer, *Characterizing Crosstalk of Superconducting Transmon Processors*, 10. März 2023, DPG Frühjahrstagung SAMOP, Hannover

3.3 Patentanmeldungen

keine

3.4 Im Themenfeld Quantencomputing geschulte Personen

Im Themenfeld Quantencomputing geschulte Personen: 5

Wie im Antrag vorgesehen gab es insgesamt 5 Teilnehmer an der von Fraunhofer IAF und IAO angebotenen „Schulungsreihe Quantencomputing“ (DHBWR: 1, EKUT: 1, UKON: 1, USTUTT: 2).

3.5 Am Projekt beteiligte Industriepartner

Insgesamt waren 5 Unternehmen am Projekt mit beteiligt, davon 2 KMU

- Bausparkasse Schwäbisch Hall AG
- Wüstenrot & Württembergische AG
- HQS Quantum Simulations GmbH
- JoS Quantum GmbH
- Airbus Defence and Space GmbH