# Arquitetura para sistema de computação quântica distribuída multi-QPU com particionamento de circuitos

**Waldemir Cambiucci, Regina Melo Silveira, Wilson Vicente Ruggiero**

Departamento de Engenharia de Computação e Sistemas Digitais
Universidade de São Paulo (USP) – São Paulo – SP – Brasil

waldemir.cambiucci@usp.br, regina@larc.usp.br, wilson@larc.usp.br

***Abstract.*** *There is a consensus that the distribution of quantum circuits between processing agents is a viable approach to achieve greater scalability with current hardware technologies, for noisy intermediate-scale quantum devices. Therefore, new quantum computer architectures with multiple processing units must consider additional circuit partitioning steps, with the generation of subcircuits with lower communication costs between partitions. This paper presents a modular multi-QPU quantum computer architecture, as well as results with hypergraphic circuit partitioning, as a permanent layer in future distributed quantum system architectures.*

***Resumo.*** *É consenso que a distribuição de circuitos quânticos entre agentes de processamento é uma abordagem viável para se obter maior escalabilidade com as tecnologias de hardware atuais, de dispositivos ruidosos de escala intermediária. Assim, novas arquitetura de computadores quânticos com múltiplas unidades de processamento devem considerar etapas adicionais de particionamento de circuitos, com a geração de subcircuitos com menores custos de comunicação entre partições. Este artigo apresenta uma arquitetura modular de computador quântico multi-QPU, assim como resultados com o particionamento hipergráfico de circuitos, como uma camada permanente em futuras arquiteturas de sistemas quânticos distribuídos.*

## 1. Introdução

A área da computação quântica promete benefícios sem precedentes na aceleração de soluções para problemas complexos, ao utilizar princípios da mecânica quântica. Entretanto, muitos desafios impedem que tais vantagens se apliquem em todos os cenários. Esses desafios estão presentes em toda a pilha de desenvolvimento de máquinas quânticas [BANDIC 2022], da implementação física de qubits ao hardware de controle [RIESEBOS 2022], do software de controle às aplicações em cenários reais. Esse retrato define a limitação do poder computacional das máquinas atuais, de dispositivos ruidosos de escala intermediária, ou *NISQ – Noise Intermediate Scale Quantum* [PRESKILL 2018], com poucos qubits, elevadas taxas de erros, latência de operações, baixa fidelidade de portas e curta duração de estados, o que impede a execução de grandes algoritmos. Essa realidade deve persistir até que computadores quânticos tolerantes a erros se tornem uma realidade.

A distribuição de circuitos entre agentes de processamento é uma abordagem viável para se alcançar maior escalabilidade com as tecnologias de hardware atuais, ao custo de um maior controle sobre a comunicação entre agentes envolvidos [DIADAMO 2021]. Arquiteturas de computadores quânticos NISQ com múltiplas unidades de processamento devem ser consideradas com etapas adicionais de particionamento de circuitos quânticos, viabilizando a geração de subcircuitos com menores custos de comunicação entre partições. Na literatura, encontramos trabalhos que exploram diferentes abordagens para sistemas distribuídos como [MONROE 2014], [LOKE 2022], assim como discussões sobre a distribuição de circuitos [YIMSIRIWATTANA 2004] e o particionamento de circuitos, através da geração de segmentos menores para distribuição [DAVARZANI 2020], [DAEI 2020]. Esses trabalhos pavimentaram a área de corte de circuitos, traçando diferentes abordagens para o desafio de segmentação de circuitos quânticos. Entretanto, uma necessidade ainda presente é a consolidação dessas técnicas no contexo de uma arquitetura modular de computadores quânticos, levando em conta os custos operacionais de comunicação, latência, acoplamento do circuito e dedicação de qubits para a comunicação entre partições. Esses são desafios reais presentes em máquinas NISQ de diferentes tecnologias para implementação de qubits físicos atuais.

Este artigo completa os experimentos sobre particionamento hipergráfico discutidos em [CAMBIUCCI 2023], apresentando uma arquitetura modular para sistemas quânticos NISQ com múltiplas unidades de processamento, com etapas para o corte de circuitos quânticos em sistemas distribuídos. Uma abordagem hipergráfica para o processo de particionamento foi adotada, com o objetivo de se obter uma distribuição mais eficiente, reduzindo custos de comunicação entre partições e permitindo um nível mais alto de abstração de circuitos, quando comparada com abordagens de trabalhos anteriores. O artigo está assim organizado: na seção 2 veremos uma breve descrição das camadas de uma arquitetura monolítica; a seção 3 apresenta a proposta de uma arquitetura de computador quântico de múltiplas QPUs, com o particionamento de circuitos com heurística hipergráfica; a seção 4 apresenta aspectos do particionamento hipergráfico e resultado de experimentos com circuitos de benchmark; o artigo finaliza com conclusões e referências.

## 2. Camadas de uma arquitetura quântica monolítica

Encontramos na literatura diferentes layouts para os componentes de uma arquitetura modular de computadores quânticos. Alguns trabalhos como [JONES 2012], [BERTELS 2020] e [BHARTI 2021] exploram etapas importantes para o tratamento e processamento de algoritmos quânticos. Em [BANDIC 2022] temos uma descrição abrangente de camadas envolvidas no processamento de computadores quânticos NISQ e uma segmentação entre compilação e o conjunto de instruções da máquina quântica, ou *ISA – Instruction Set Architecture*. [JONES 2012] faz um destaque para a camada de correção de erros quânticos, determinante para a computação quântica tolerante a erros. Atualmente, essa camada de correção de erros quânticos está migrando para o hardware clássico de controle em algumas arquiteturas e plataformas [RIESEBOS 2022]. Em [FU 2017] encontramos aspectos relevantes para a construção de uma microarquitetura de computadores quânticos, com destaque para a conversão entre a representação de alto nível e as linhas de microinstruções, responsáveis pelas transformações sobre qubits na camada física de hardware.

A partir dessa literatura, podemos descrever uma arquitetura monolítica através de três camadas principais, a seguir:

**Camada de programação:** onde encontramos os recursos de codificação, bibliotecas e frameworks disponíveis na plataforma, para o desenvolvimento de programas quânticos. Atualmente, cada fornecedor de plataforma de mercado, como IBM, MICROSOFT, GOOGLE ou RIGETTI oferece um conjunto amplo de recursos para o desenvolvedor, como ferramentas de programação, simulação, estimativa de recursos e visualização de resultados.

**Camada de compilação:** a partir da descrição do circuito em linguagem de programação de alto nível, nessa camada o circuito sofre transformações que o tornam compatível com o processador alvo para execução. Podemos relacionar diferentes etapas relacionadas como decomposição, otimização, agendamento, mapeamento e síntese. Essas atividades manipulam as portas e operações quânticas do circuito, com diversos objetivos, como: - preparação de estados quânticos e inicialização eficiente de qubits; - mapeamento físico de qubits de acordo com a conectividade no hardware; - agrupamento de operações por equivalência e remoção de redundâncias; - tradução de operações nas portas básicas suportadas pelo processador alvo; - decomposição e síntese de circuitos quânticos; - e otimização de circuitos em função de hardware alvo.

**Camada de execução:** a partir da síntese do circuito quântico otimizado, realizamos a tradução das instruções em comandos da eletrônica envolvida no controle do hardware quântico e qubits físicos. Atualmente, muitas tarefas de preparação, controle e medição estão movendo para o hardware de controle clássico assim como novas definições para a microarquitetura quântica em uso [FU 2017].

## 3. Arquitetura de computador quântico de múltiplas QPUs

Considerando as necessidades para o particionamento de circuito e distribuição eficiente de cargas de processamento, propomos a arquitetura de referência para computadores quânticos de múltiplas unidades de processamento na figura 1, com etapas de preparação e execução de particionamento de circuitos.

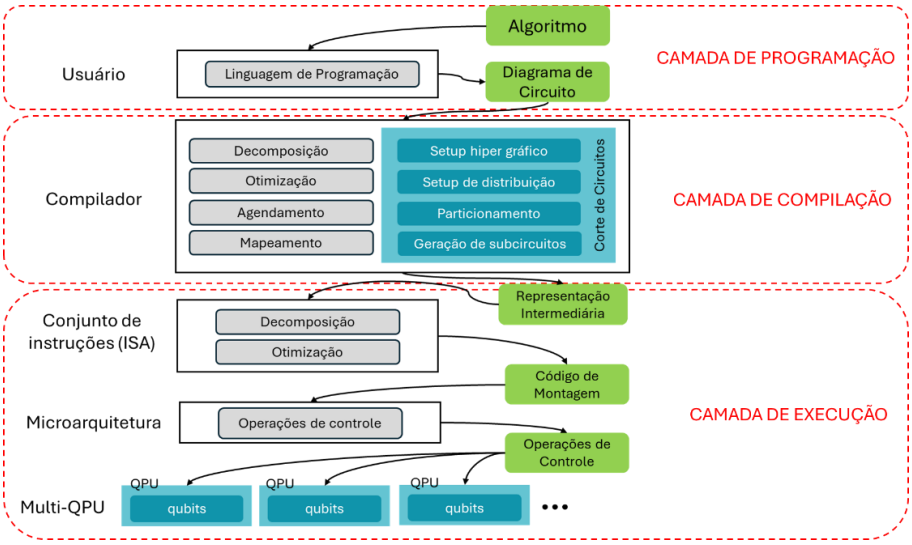


**Figura 1. Arquitetura de referência para computador quântico com múltiplas unidades de processamento e etapas de particionamento hipergráfico para distribuição eficiente de subcircuitos. Fonte: elaborado pelo autor [2024].**

No desenho proposto, é possível identificar as camadas de programação, compilação e execução, assim como as etapas adicionais para preparação e corte de circuitos. Como desafio no processo de distribuição de circuitos, precisamos tratar a função objetiva relacionada com o custo de comunicação entre partições. O objetivo é reduzir o número de ebits ou qubits dedicados para a comunicação entre partições, ao longo da execução do circuito, assim como otimizar o uso de enlaces entre subpartições ao longo da execução [MARTINEZ 2019] e [DAVARZANI 2020].

## 4. Particionamento hipergráfico de circuitos

Como estratégia utilizada no particionamento de circuitos exploramos a representação hipergráfica do circuito de entrada, para posterior uso de heurísticas de particionamento em duas configurações: a primeira é baseada em **heurística de particionamento hipergráfico ponderado**, utilizando a razão de acoplamento do circuito como peso em hiperarestas; a segunda configuração considera **as restrições de conectividade entre qubits** presentes nas topologias de máquinas alvo, gerando pesos em hiperarestas do hipergrafo. Como métrica de impacto no circuito, a razão de acoplamento pode ser calculada a partir do número de portas quânticas entre múltiplos qubits, sobre o total de qubits presentes no circuito considerado. Essa medida faz uma indicação sobre o custo de particionamento do circuito: quanto maior a razão de acoplamento, maior o custo potencial gerado pelo corte do circuito, o que exige heurísticas otimizadas para particionamentos mais eficientes.

Inúmeras heurísticas de particionamento são exploradas na literatura, como KL [KERNIGHAN-LIN 1970] e SW [STOER-WAGNER 1997]. Focamos a heurística de FM [FIDUCCIA-MATTHEYSES 1982] como base para o processo de particionamento hipergráfico, devido sua implementação simples e reduzido tempo de execução, com uma função Big-O ( n ), isto é, de tempo linear. Sua escolha foi motivada pelos resultados discutidos no trabalho de [CAMBIUCCI 2023], a partir da comparação de diferentes heurísticas de particionamento gráfico e hipergráfico no contexto de circuitos quânticos.

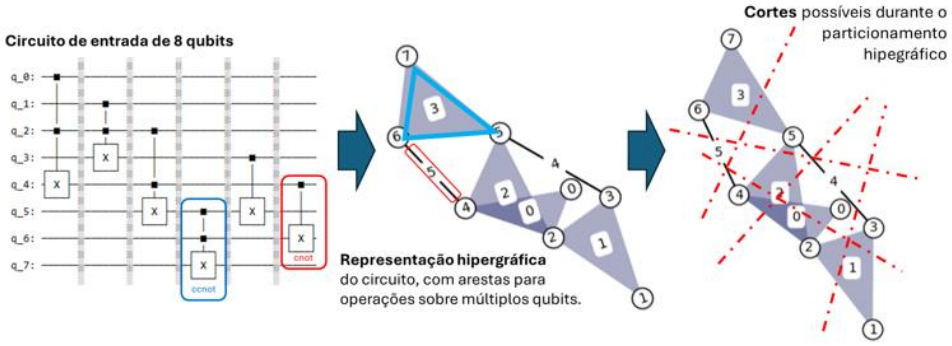


**Figura 2. Representação hipergráfica de circuitos quânticos e possíveis cortes durante o processo de particionamento. Fonte: elaborado pelo autor [2024].**

A heurística de FM é um método eficaz para dividir um conjunto de vértices de um hipergrafo em dois grupos ou partições, minimizando as relações entre eles. O algoritmo funciona em etapas iterativas, onde cada vértice é movido de uma partição para outra, se isso reduzir o número de arestas entre as subpartições. Criamos uma variação desse algoritmo, onde uma telemetria ponderada é aplicada para guiar o processo de movimento dos vértices. Essa telemetria considera pesos nas hiperarestas, com base no número de conexão entre um vértice e uma partição. Os pesos são atualizados dinamicamente ao

longo das iterações, permitindo ao algoritmo priorizar movimentos que resultam em uma redução significativa de comunicação entre partições. Ao considerar não apenas a estrutura do circuito, mas também pesos nas hiperarestas, o algoritmo é capaz de alcançar uma divisão mais eficiente, reduzindo os custos de comunicação entre partições para sistemas distribuidos. O particionamento bipartido centralizado foi abordado por ser amplamento utilizado como teste de calibração, sendo realizado pelo corte central na largura do circuito de entrada, ou seja, cada uma das duas partições tem mesmo número de qubits. A tabela 1 ilustra os resultados desse experimento em destaque:

**Tabela 1. Bipartição para circuitos de n qubits e profundidade d, variando num. de cnots; bipart é o total de ebits entre partições; fm é o total de ebits obtido com Fiduccia-Mattheyses; red.% a redução de comunicação obtida com heurística hipergráfica.**

| Circuit | n | d | cnots | bipart | fm | red.% | Circuit | n | d | cnots | bipart | fm | red.% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **QFT** | 4 | 31 | 18 | 5 | 4 | 20% | **QPE** | 4 | 28 | 11 | 4 | 2 | 50% |
| **QFT** | 6 | 47 | 39 | 12 | 9 | 25% | **QPE** | 6 | 54 | 34 | 12 | 8 | 33% |
| **QFT** | 8 | 63 | 68 | 22 | 16 | 27% | **QPE** | 8 | 84 | 65 | 22 | 16 | 27% |
| **QFT** | 10 | 79 | 105 | 35 | 25 | 29% | **QPE** | 10 | 94 | 98 | 35 | 24 | 31% |
| **QFT** | 15 | 119 | 231 | 77 | 56 | 27% | **QPE** | 15 | 168 | 231 | 77 | 56 | 27% |
| **QFT** | 30 | 239 | 915 | 330 | 225 | 32% | **QPE** | 30 | 342 | 910 | 330 | 225 | 32% |
| **QFT** | 50 | 399 | 2525 | 925 | 625 | 32% | **QPE** | 50 | 588 | 2522 | 925 | 625 | 32% |
| **QFT** | 100 | 799 | 10050 | 3725 | 2500 | 33% | **QPE** | 100 | 1188 | 10047 | 3725 | 2500 | 33% |
| **QFT** | 120 | 959 | 14460 | 5370 | 3600 | 33% | **QPE** | 120 | 1428 | 14457 | 5370 | 3600 | 33% |

Com a abordagem hipergráfica usada, a redução de cortes em portas binárias ficou em torno de 35% em relação ao particionamento central randômico, quando simplesmente cortamos o circuito em sua largura sem reordenamento de qubits. Foram utilizados benchmarks de Quantum Fourier transform (QFT) e Quantum Phase Estimation (QPE), gerados via ferramenta MQTBench [QUETSCHLICH 2023]. A máquina utilizada foi um notebook de 16GB RAM com 2,20GHz com 12 cores e 16 processadores.

## 5. Conclusões

Este artigo propõe um desenho modular de arquitetura de computador quântico multi-QPU, com uma abordagem hipergráfica baseada em heurística de FIDUCCIA-MATTHEYSES para otimizar o particionamento de circuitos, reduzindo os custos de comunicação entre partições, como discutido anteriormente em [CAMBIUCCI 2023].

## Referências


BANDIC, Medina; FELD, Sebastian; ALMUDEVER, Carmen G. Full-stack quantum computing systems in the NISQ era: algorithm-driven and hardware-aware compilation techniques. In: 2022 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2022. p. 1-6.

BERTELS, K. O. E. N. et al. Quantum computer architecture: Towards full-stack quantum accelerators. In: 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2020. p. 1-6.

BHARTI, Kishor et al. Noisy intermediate-scale quantum (NISQ) algorithms. arXiv preprint arXiv:2101.08448, 2021.

Brian W KERNIGHAN and Shen LIN. An efficient heuristic procedure for partitioning graphs. The Bell system technical journal, 49(2):291–307, 1970.

CAMBIUCCI, W., SILVEIRA, R. M., and RUGGIERO, W. V., "Hypergraphic Partitioning of Quantum Circuits for Distributed Quantum Computing," 2023 IEEE International Conference on Quantum Computing and Engineering (QCE), Bellevue, WA, USA, 2023, pp. 268-269, doi: 10.1109/QCE57702.2023.10237.

DAEI, Omid; NAVI, Keivan; ZOMORODI-MOGHADAM, Mariam. Optimized quantum circuit partitioning. International Journal of Theoretical Physics, v. 59, n. 12, p. 3804-3820, 2020, https://arxiv.org/abs/2005.11614

DAVARZANI, Zohreh et al. A dynamic programming approach for distributing quantum circuits by bipartite graphs. Quantum Information Processing, v. 19, n. 10, p. 1-18, 2020. https://arxiv.org/abs/2005.01052

DIADAMO, Stephen; GHIBAUDI, Marco; CRUISE, James. Distributed quantum computing and network control for accelerated VQE. arXiv preprint arXiv:2101.02504, 2021.

FIDUCCIA,C.M., MATTHEYSES, R. M., "A Linear-Time Heuristic for Improving Network Partitions," 19th Design Automation Conference, 1982, pp. 175-181, doi: 10.1109/DAC.1982.1585498.

FU, Xiang et al. An experimental microarchitecture for a superconducting quantum processor. In: Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture. 2017. p. 813-825.

JONES, N. Cody et al. Layered architecture for quantum computing. Physical Review X, v. 2, n. 3, p. 031007, 2012.

LOKE, Seng W. From Distributed Quantum Computing to Quantum Internet Computing: an Overview. arXiv preprint arXiv:2208.10127, 2022.

MARTINEZ, Pablo; HEUNEN, Chris. Automated distribution of quantum circuits via hypergraph partitioning. Physical Review A, v. 100, n. 3, p. 032308, 2019.

MONROE, C. et al. Large-scale modular quantum-computer architecture with atomic memory and photonic interconnects. Physical Review A, v. 89, n. 2, 2014.

PRESKILL, John. Quantum computing in the NISQ era and beyond. Quantum, v. 2, p. 79, 2018. https://arxiv.org/abs/1801.00862

QUETSCHLICH, Nils; BURGHOLZER, Lukas; WILLE, Robert. MQT Bench: Benchmarking software and design automation tools for quantum computing. Quantum, v. 7, p. 1062, 2023.

RIESEBOS, Leon et al. Modular software for real-time quantum control systems. In: 2022 IEEE International Conference on Quantum Computing and Engineering (QCE). IEEE, 2022. p. 545-555.

STOER, M., WAGNER, F. (1997). "A Simple Min-Cut Algorithm." Journal of the ACM, 44(4), 585-591.

YIMSIRIWATTANA, Anocha; LOMONACO JR, Samuel J. Distributed quantum computing: A distributed Shor algorithm. In: Quantum Information and Computation II. SPIE, 2004. p. 360-372. https://arxiv.org/abs/quant-ph/0403146.