QAMP 2021 : Operationalizing Quantum Kernels

Mentees : Cheryl Fillekes & Michaël Rollin

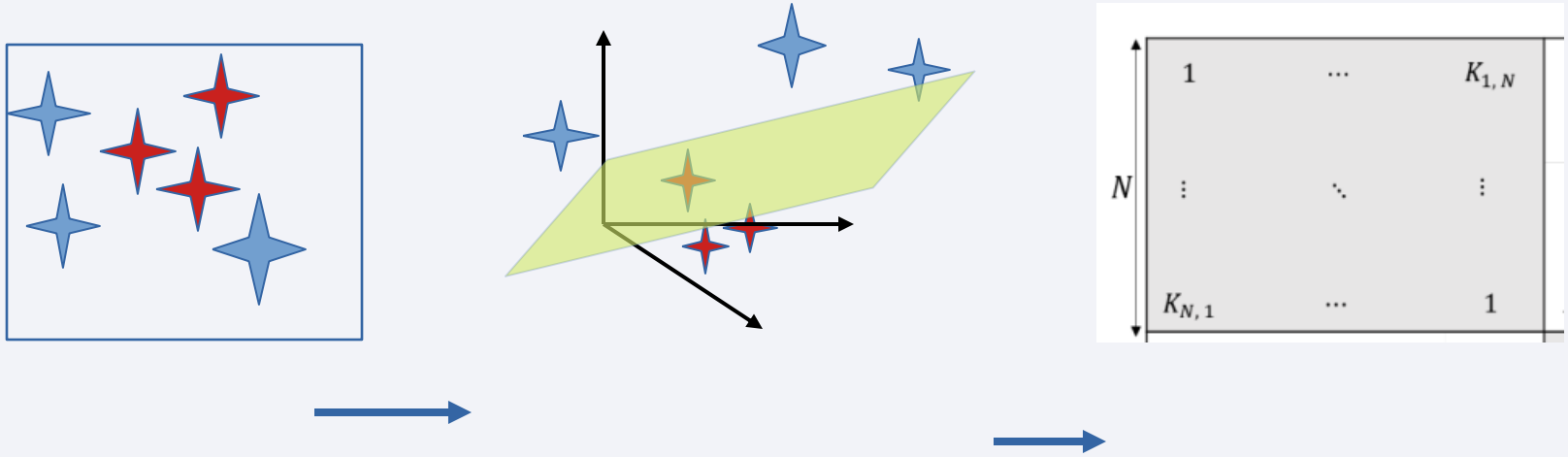
Mentor : Travis Scholten

Voice : Neural network cloned voice of Michaël Rollin

The project

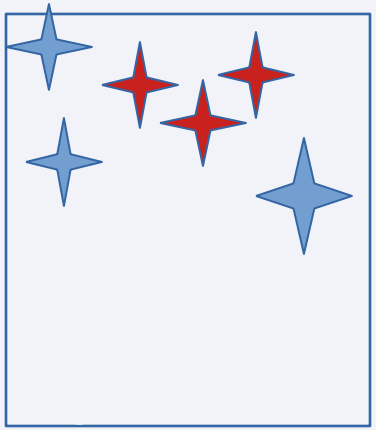
Implement algorithms in Naveh, Fitzgerald, Phan, Lockwood & Scholten 2021 ArXiv:2112:08449v1 **Kernel Matrix Completion for Offline Quantum-Enhanced Machine Learning**

Architect processing of quantum **Machine Learning kernels** to streaming data workload.

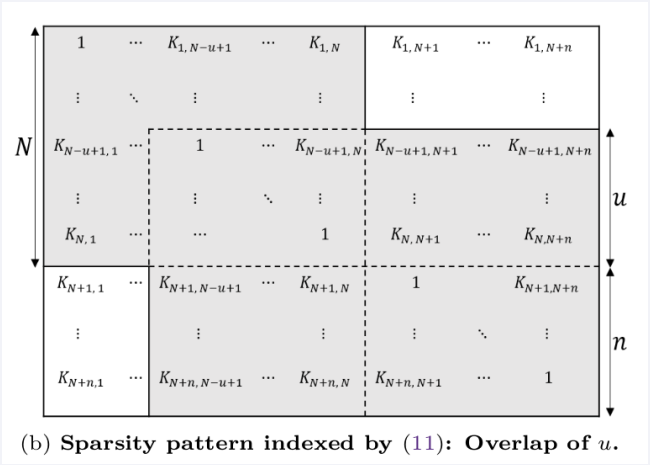


Goal : Determine the impact of different parameter choices on reconstruction of quantum Machine Learning kernel in real time, e.g.:

- number of qubits
- depth of circuits
- size of data blocks
- volume of data stream

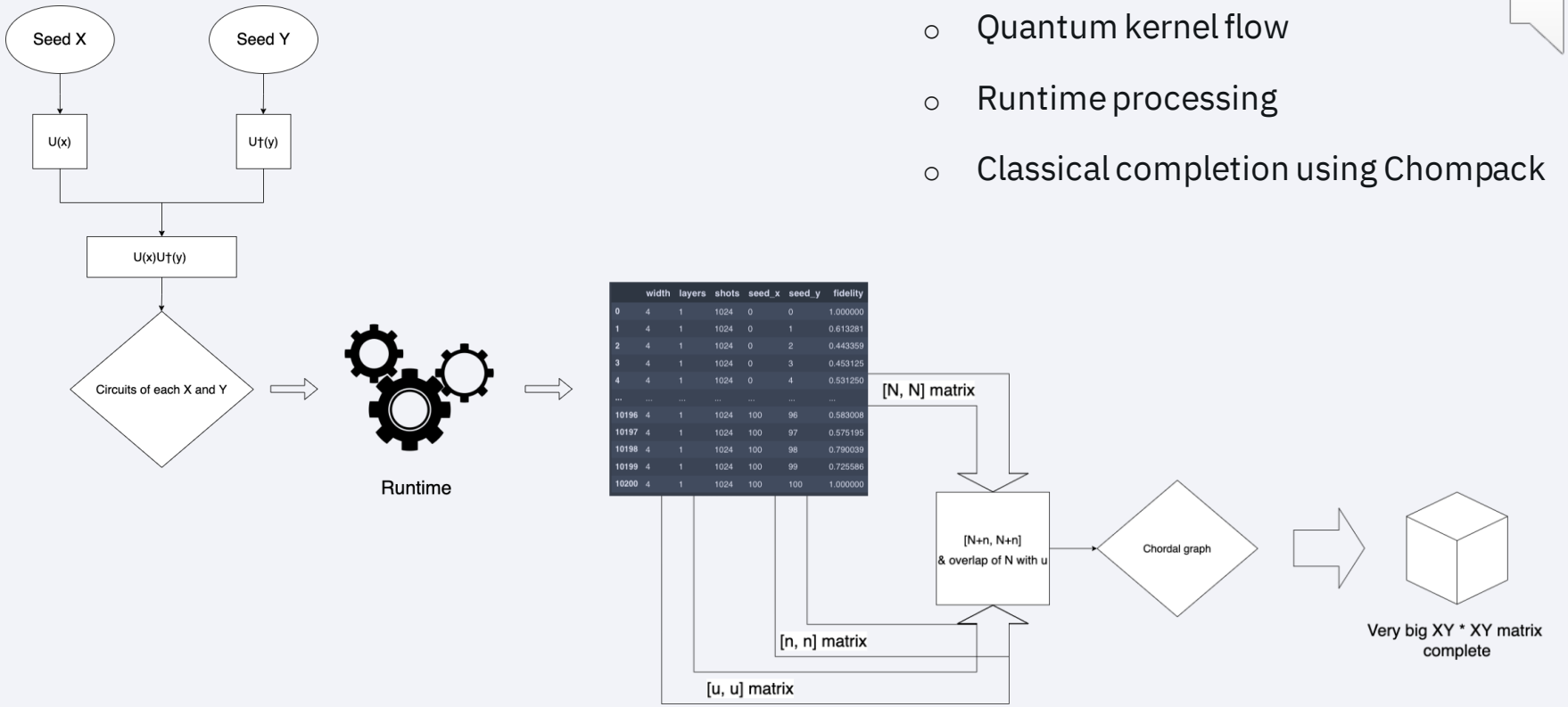


What are the limits on these data rates for different kinds of Machine Learning kernels?



(b) **Sparsity pattern indexed by (11): Overlap of u.**

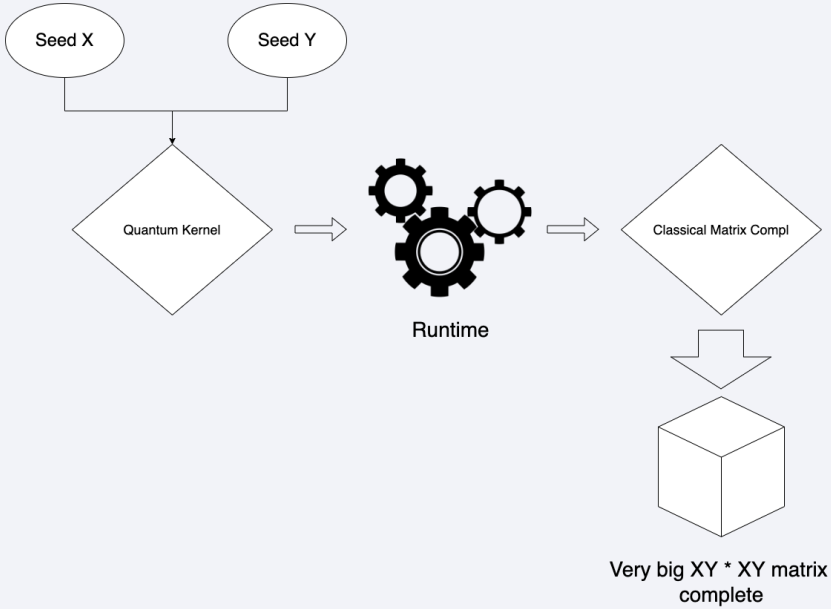
The workflow



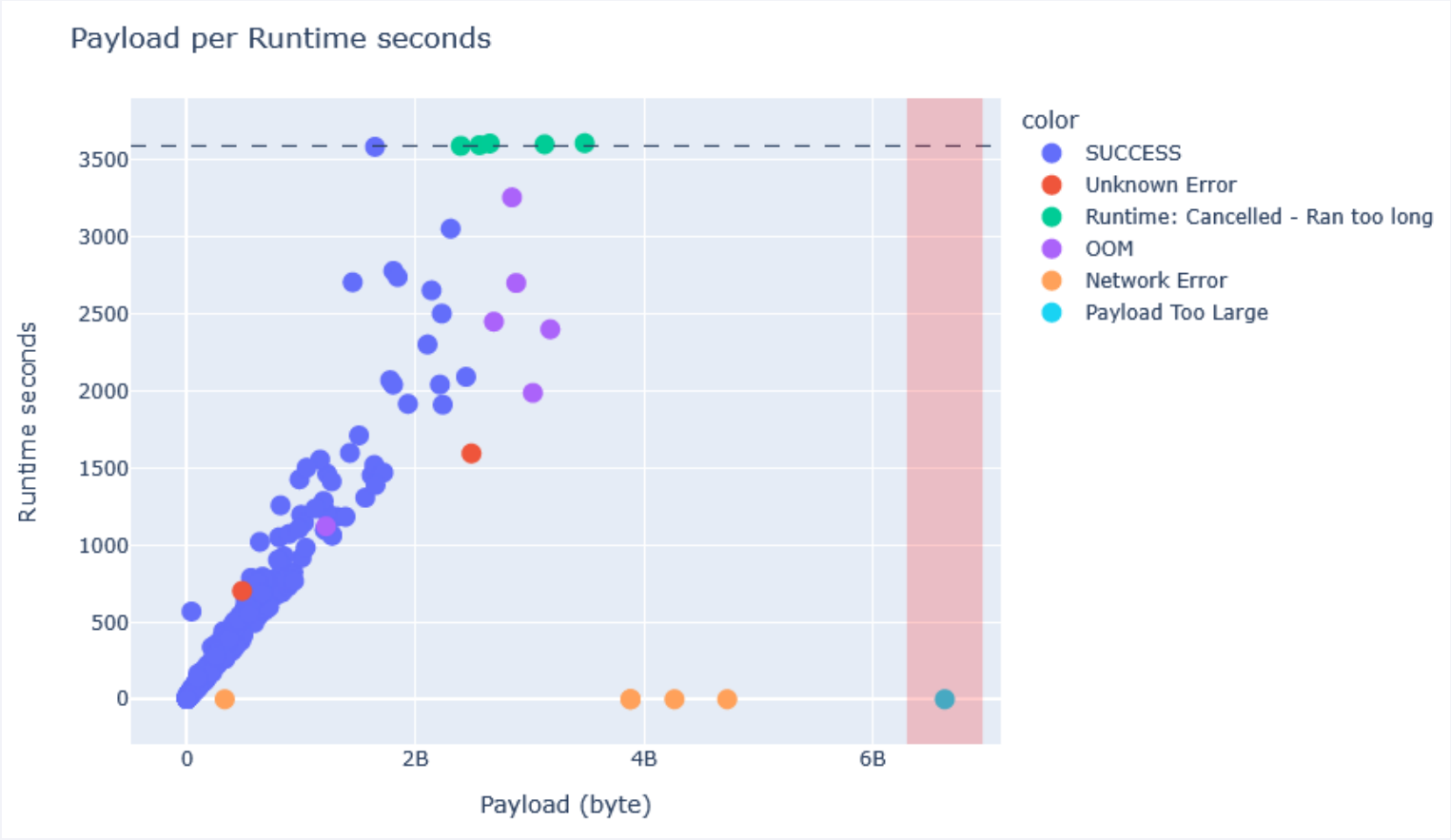
- Quantum kernel flow
- Runtime processing
- Classical completion using Chompack

Functional structure

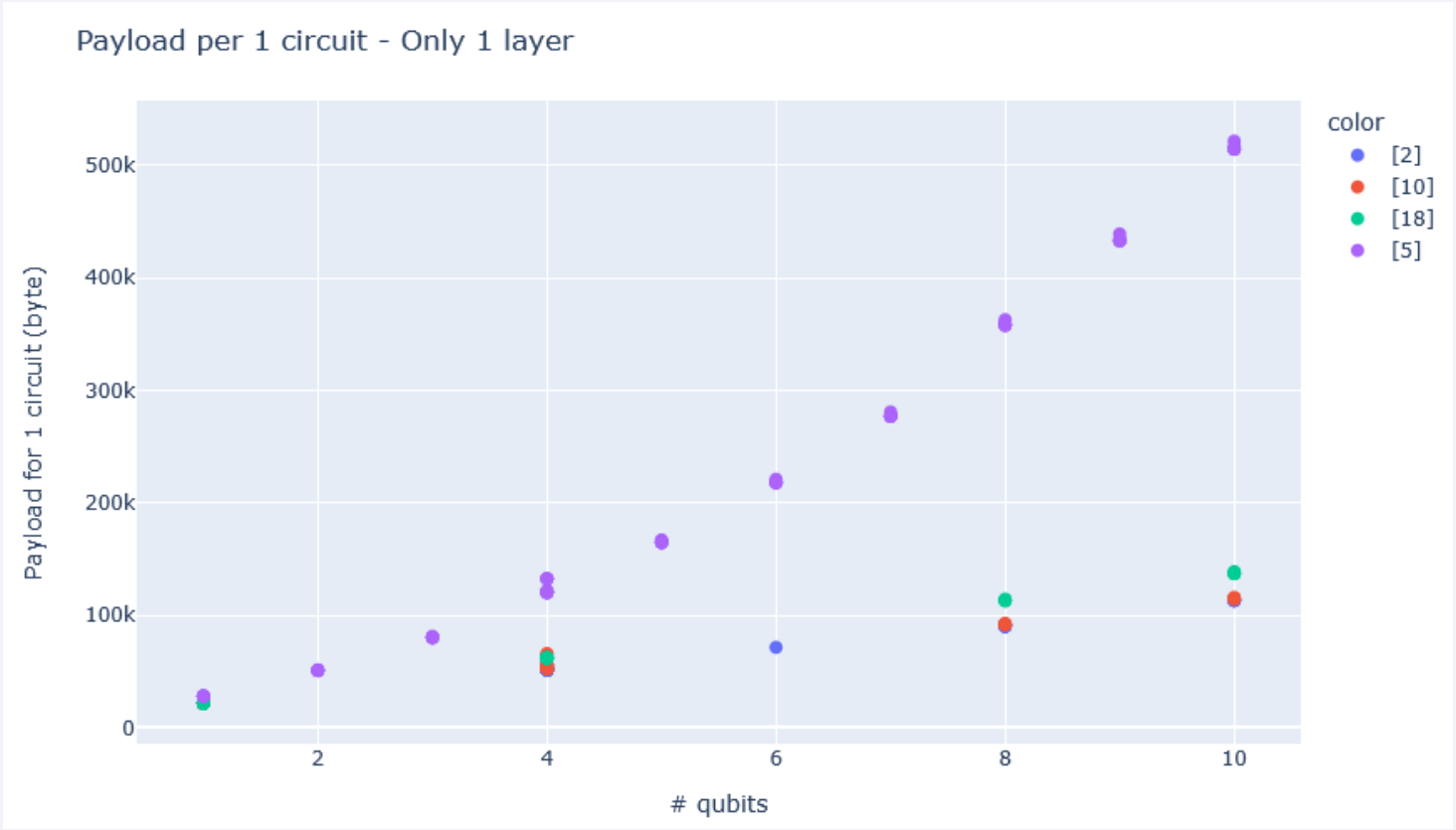
- Inputs :
  - Kernel matrix size
  - Circuit template
  - Number of qubits/layers
  - Backend
- Telemetry from Runtime :
  - Processing time
  - Number of circuits sent
  - Payload size
  - Error



Runtime seconds



Payload information

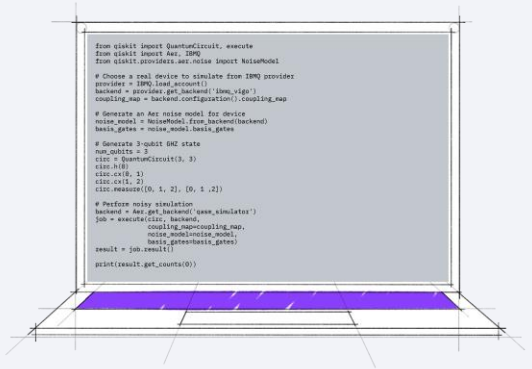


Recap

- Our goals :
  - Quantum kernel program ✅
  - Automation ✅
  - Mainframe integration ✅
  - Runtime study ✅
  - Matrix completion ✅
  - Workflow end-to-end endpoint ✕
  - Matrix compl study 🔄
  - Containerization on IBM System Z* ✅

*aka "IBM Mainframe" or "LinuxONE"

- Future
  - Documentation 🔄
  - Adapt kernel program
  - Python package
  - Add project to Qiskit Ecosystem
  - Optimize Classical Part on AIA, DLC, TPU, GPU
  - Port Qiskit Runtime to IBM System Z



Faster Quantum/Classical Hybrid ML Algorithms: IBM z16 Telum Chip + Quantum Processing

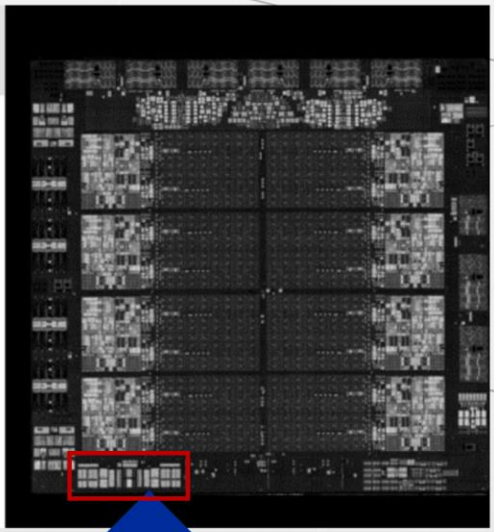
The IBM Telum Processor Design

Performance and Scale
- Optimized core
- New cache hierarchy & multi-chip fabric

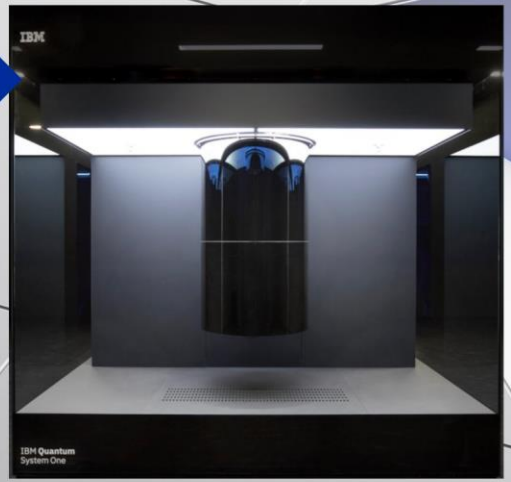
Embedded Accelerators
- Sort, Compression, Crypto
- AI

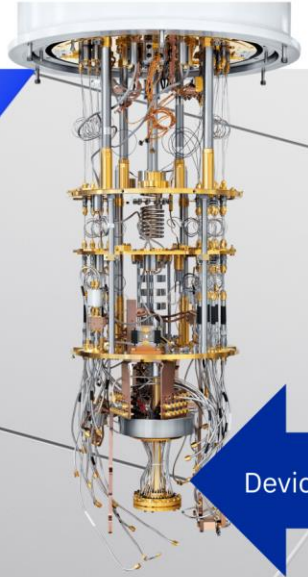
Industry-leading Security
- Encrypted Memory
- Improved Trusted Execution Environment

Unmatched Reliability and Availability
- L2 cache SRAM wipe-out error correction & sparing
- 8-DIMM Redundant Array of Memory (RAIM)



REST API

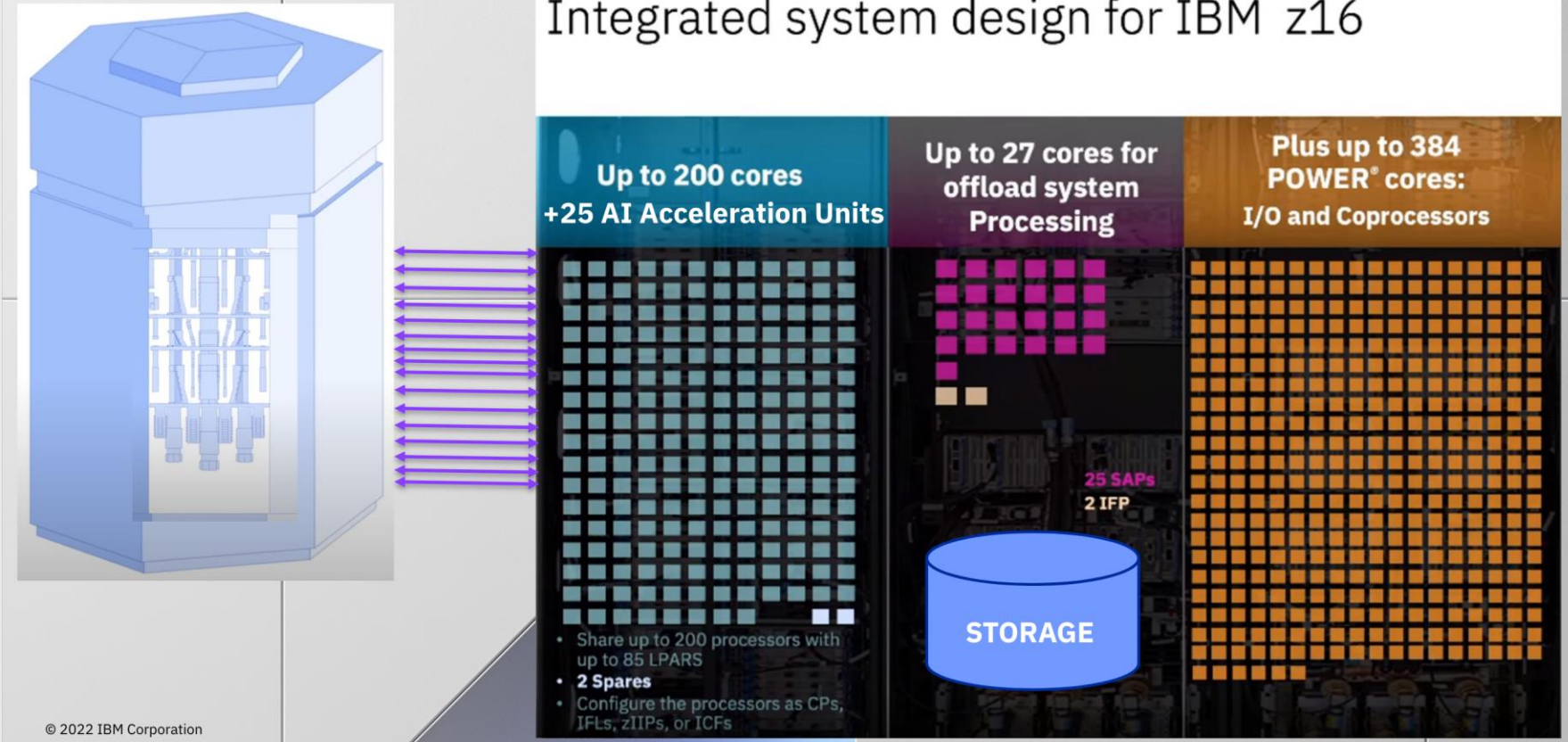




Device Interface

IBM z16 AI Accelerator shares cache with 8 CPUs

Process Data *in-place* and *in parallel*



Thanks for your attention !

Cheryl, Michaël & Travis