## Resource Management & Scheduling

### **Goal: Explore current vendor &lab approaches to hybrid scheduling and collect key requirements.**

### **Session Chairs**

* **ORNL:** Thomas Naughton
* **IBM:** Iskandar Sitdikov

### **Prompts**

1. **What are the highest priority requirements for resource management & scheduling?**
2. Which hybrid workflows (e.g. VQE, QAOA) do vendors prioritize, and why?
3. Are there hybrid workflows which would benefit from simultaneous (lockstep) vs interleaved (time‑slice) QPU allocations?
   1. What is the timescale of quantum operations that would justify simultaneous vs interleaved allocation?
   2. What are the classical/quantum resource requirements and how will they dictate resource allocation?
4. What are the perceived trade‑offs of single‑level (flat) vs hierarchical (two‑level) scheduling in practice?
5. What metrics need to gathered/exposed from the quantum resources to the resource management system?
   1. What operational data should quantum vendors provide to classical operations?
6. How should we delineate responsibilities between Workload Manager Systems (e.g. Pilot, Nextflow) and the Quantum Software Stack?
   1. Which interfaces or plugins must the QSS expose to integrate seamlessly with the WMS?
7. What level of abstraction should quantum resources be represented as? qubit? node? topology?
8. What data‑motion strategies (staging, streaming, caching, recovery) are needed for the hybrid applications?
   1. Are there unique elements for quantum / hybrid scenarios?
9. (Optional) Is there a need for dynamic discovery of system capabilities as relates to resource management & scheduling?
10. Next Steps, Open Questions?, Key Challenges (near/far)

### **Summary Slides Outline**

* **Motivation**
* **Current Status**
* **Challenges (near/long term)**
* **Open Questions**
* **Next Steps**
* **Overlap**

### **Notes from Breakout Groups**

#### **Breakout Group 1**

Participants:

* a

Notes

* Minimal reboot time
  + Minimal switch time
* Uptimes of the systems
  + Ways to fix it
    - Being prompt with users
    - Having backup capacity
  + Different systems have different capabilities if you want to go for backup
* Device sharing
  + What is a resource?
    - Entire device vs qubits
    - 8 / 15
    - Can we with software design this interleaved system
  + What vendors want
  + Probably we do not need sharing of devices
* Number of qubits per usecase is a question of devices available or application?
* Profiling / metrics
  + GPU hours, Node hours. Maybe QPU hours is metrics.
  + Calibration schedules and calibration data
  + Any metrics should be exposed to users
    - Papers are often requires that
  + Environmental data
* Maybe allocation time is an open-question
* How quantum resources are attached to compute infrastructure: notes vs shared.
  + Near term: separate partition that has access to QPU
* Should we have interleaved execution of shots?
* How to run jobs close to calibration? Needs specification
  + Maybe high priority based on price
  + Some users want to have calibration running now
  + Granular access to calibration (full, partial, etc.).

#### **Breakout Group 2**

Participants:

* a

Notes

* Maximize resource usage vs isolation of resources
  + Pulse level vs application developers
  + Sharing resources
  + Why we see usecases on smaller amount of qubits
  + For single users reuse of full chip is making sense
  + Paper for Georgia tech
* Availability of devices
  + Scarcity of resources
  + **Maybe Quantum should drive scheduling, not HPC**
* Applications
  + There is a huge separation between depth + width of circuits.
    - We want to make noise of machine stable
    - Multiple ports to different machines
    - Scheduling and noise levels are not stable
* Interleaved vs full allocation
  + As we scaling up loading and uploading will be challenging
    - How scheduling is affected by that
* What is job?
  + Such an interesting question
* Do you need to have conditional data readout?
* Complete description of hybrid job?
  + We should look at what has been done with ML world
    - MLIR dialects
* **Follow up to estimate of MLIR payload sizes**
* Resource utilization
  + Idle time

#### **Breakout Group 3**

Participants:

* a

Notes

* Levels of abstraction and levels of programmability
  + My whole hybrid job is set and forget
  + Ping-pong workflow
  + Users want to do simultaneous intra-kernel communication
* How much users are using?
  + How much resources are allocated?
    - This affects grants and cost structure
  + Partition project based on intent
* Quantum resource estimation
* Time allocation vs quality of results
* Resources availability
* Benchmarking across qc architectures
* As an application developer I want a scheduler to take care of resources when I need them. I want to have some options to be hidden but still have an ability to control it (via flags, etc).
  + We need to have Slurm package that works for both quantum and classical
* What metrics do vendors need to expose?
  + Personas
    - Performance engineers
      * Access to lowest level of info
      * Mean time to failure
      * Mean time to recovery
        + Cost of recovery of instrastructure
      * Time on device
        + Idle time on device
      * Voltage scaling
    - Algorithms
      * Automatic program for profiling
* Interleaving vs not
  + Have a flag not to interleave
* Performance analysis
* How many qubits, how many gates, etc? This will affect how scheduling will be done.
  + **This is what I’m doing and give me back estimates on resources.**
* Streaming data back and forth between quantum and classical infra.

#### **Breakout Group 4**

Participants:

* a

Notes

* Queues
  + Local emulator
  + Availability of resources
  + Queue prediction
  + Waiting times are damaged by wait times
* Interleaved vs not
  + Postprocessing
  + QAQC interleaved will benefit
* Flat scheduling makes life easier
* Metrics
  + Energy drain
  + Utilization
  + Metrics consumption for mitigation and quantum compute
  + Profiling
    - Traces
* Do we have enough tools or we need something new?
* Resource estimation for quantum time

Note At Outbriefing

* Question about benchmarking, making sure the metrics and data come together. This was not focus of discussions in the resource mgmt sessions, but there is certainly overlap between benchmarking and the resource estimates you would put into the resource manager (job description).