

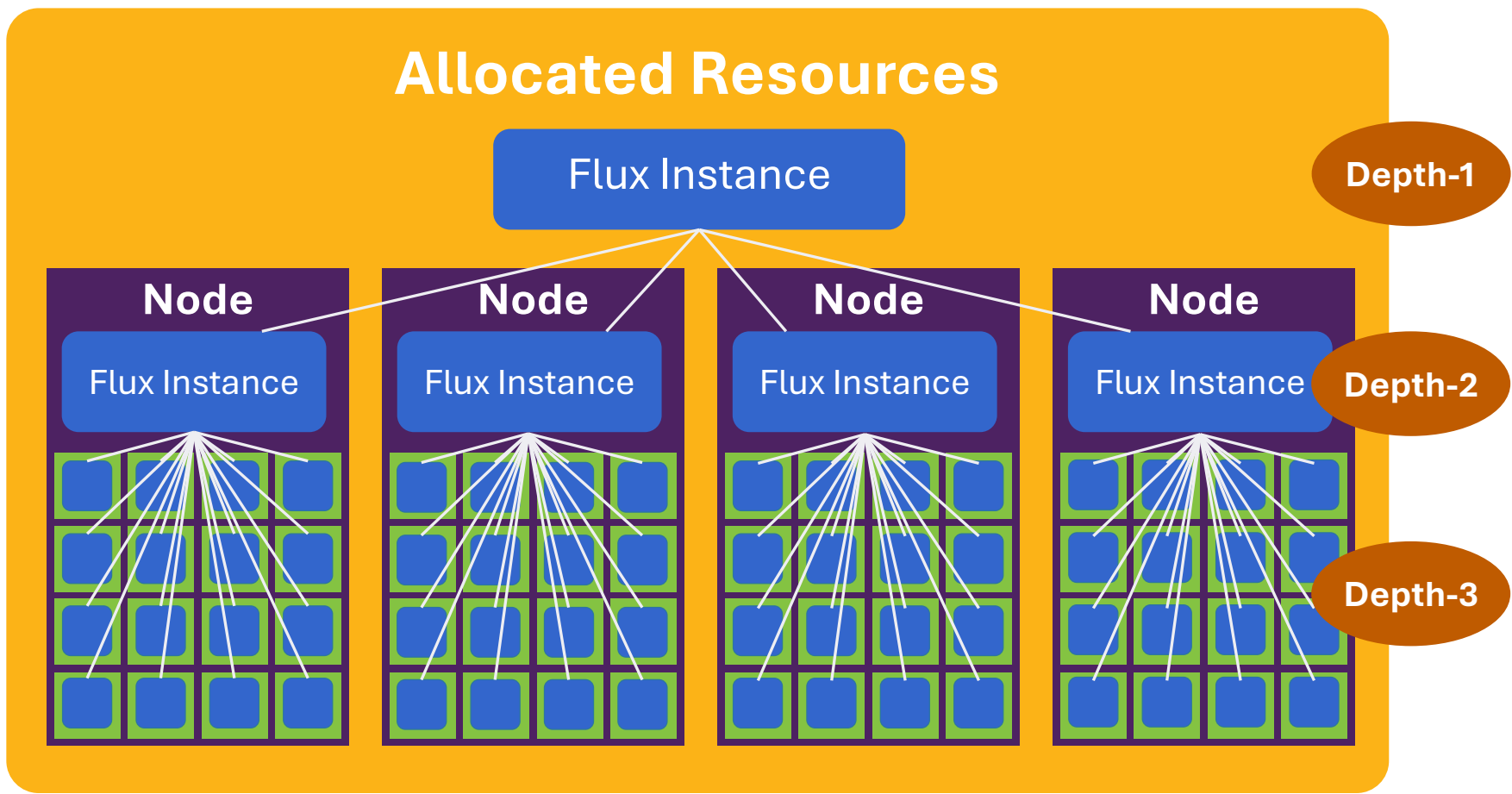
Flux: Next-Generation Resource Management at Extreme Scale

Revolutionizing workflows and I/O

James Corbett and the Flux team: A. Chu, R. Day, J. Garlick, M. Grondona, W. Hobbs, D. Milroy, Z. Morton, C. Moussa, T. Patki, B. Rountree, V. Sochat, T. Scogland, R. Springmeyer, J.-S. Yeom (LLNL)

Flux is a fully functional user- and system- level workload manager, and the framework components can be used to build custom resource managers. It was created to resolve growing issues with classical resource managers designed in the era when nodes had one socket and one core and clusters were homogeneous. In supporting **fully hierarchical scheduling**, Flux also became easy to run under other systems, making it a portable and efficient scheduler for a wide variety of workflows.

- 3–8 FTEs over 11 years, funded by the ASC program



“Sub-Flux” instances allow scheduler customization and full user control.

Production-Ready on El Capitan

- Over the past year, Flux has transitioned from running primarily on test and development clusters to running on the Advanced Technology Systems, including El Capitan, Tuolumne, and Early Access Systems. Users are running hundreds of jobs across tens of thousands of nodes every day.
- The long-term plan is for Flux to roll out as a system resource manager on all LC clusters.
- Recent work has made Flux ready for machines or cloud clusters **even larger** than El Capitan.
- Other sites are also adopting and experimenting with Flux.



Rabbits: A Novel Exascale I/O Model

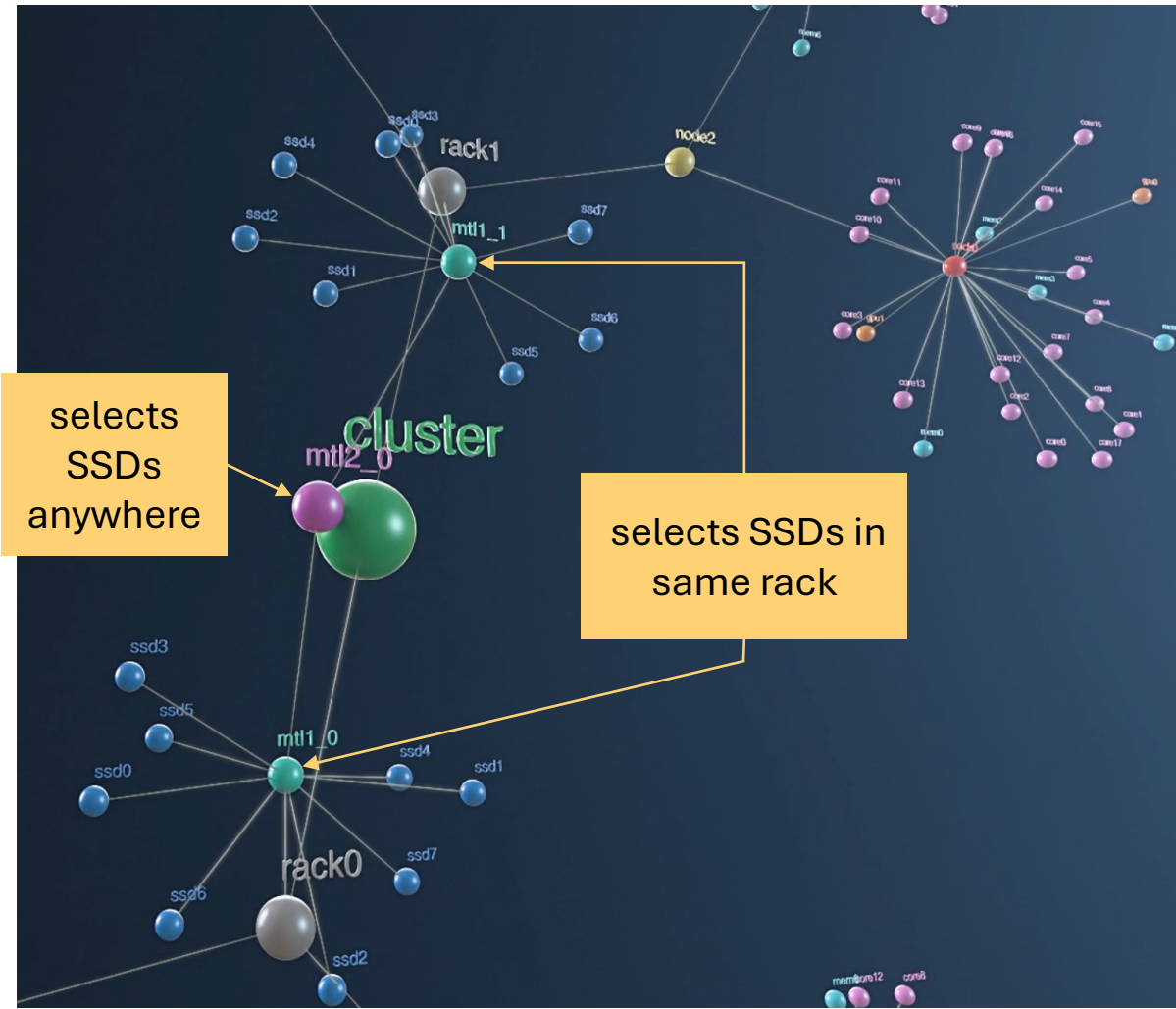
Rabbits are a first-of-its kind I/O storage solution for exascale, providing SSDs directly PCIe-attached to compute nodes, as well as network-attached SSDs for ephemeral or persistent file systems.

- Rabbit integration and support required the features and flexibility offered by Flux; infeasible with the older generation of resource managers.
- Operational on El Capitan systems and in early use.
- Through Flux, users can request several different types of dynamically configured Rabbit file systems for faster performance.



Rabbits can act as an intermediary between compute nodes and Lustre, LC's permanent file system. Communication with the Rabbits required integrating with cloud technologies such as Kubernetes.

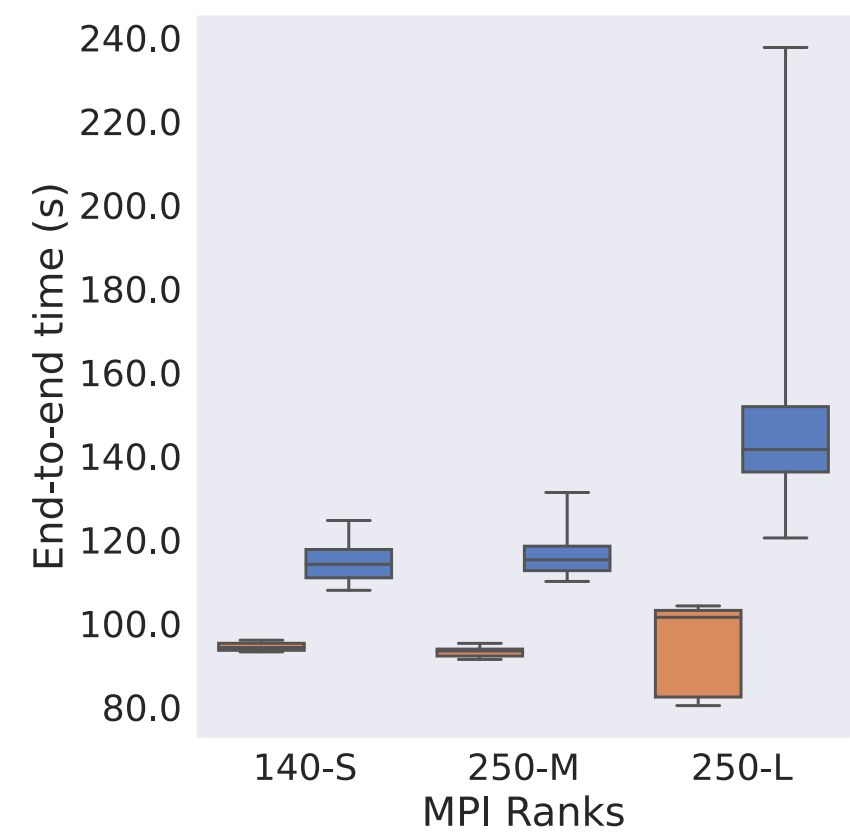
Flux pioneered directed graph-based scheduling to manage complex combinations of extremely heterogeneous resources, such as the Rabbits.



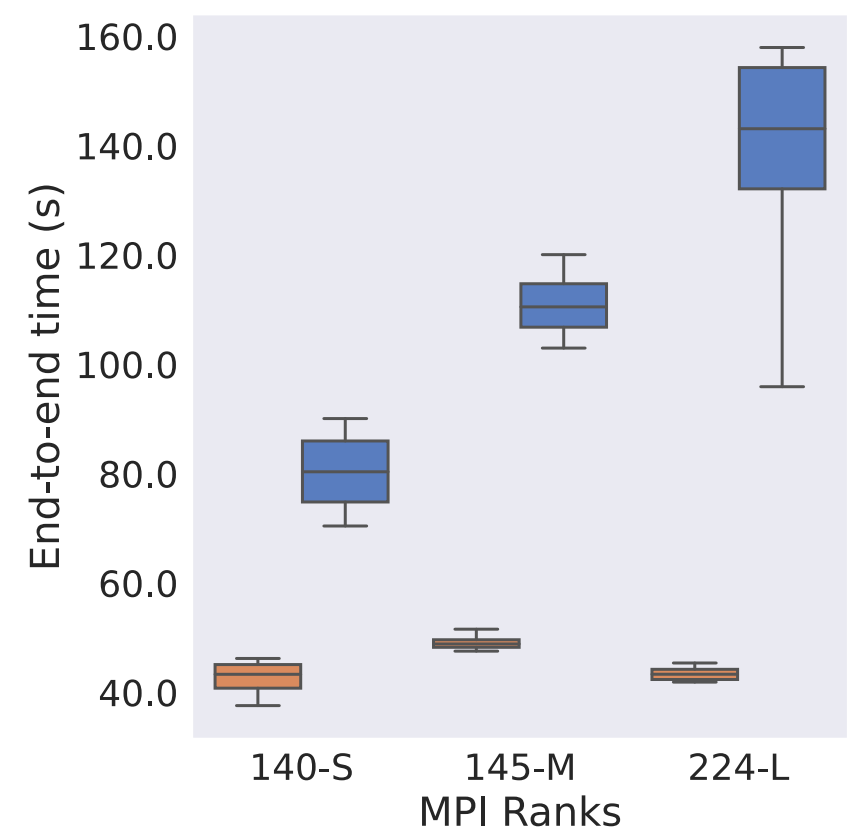
- Flux integration with Rabbits has been a 4+-year, high-bandwidth collaboration between teams at Livermore and HPE.
- Collaboration has leveraged Flux features to reduce Rabbit software system noise and provide secure data movement.

Support for Flexible Workflows in HPC and Cloud Environments

- Recent work has demonstrated the power of Flux in a cloud context.
- Flux plugin for Kubernetes schedules up to **3.5x faster** than the native scheduler, with less variability.
- In HPC clusters, jobs can now shrink on demand.



Timing for Flux (orange) vs Kubernetes native scheduler (blue) on QMCPack (left) and LAMMPS (right).



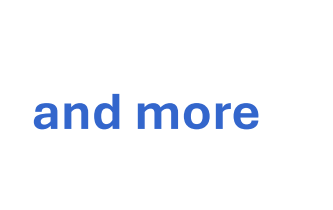
Ongoing and Future Work

- Support for power-aware scheduling
- Support for growing jobs on demand
- Support for managing workflows across multiple clusters

Conclusions

Flux supports modern workflows at extreme scale, leveraging expertise gained from developing and running RMs including SLURM (2002–present) and its predecessor, LCRM. Flux now plays a critical role in the system software stack on LC's largest machines.

Collaborators



Providing next-generation workflow tools and resource management