SAND2018-3827PE

Can RoCE Support Diverse Cluster Workloads?

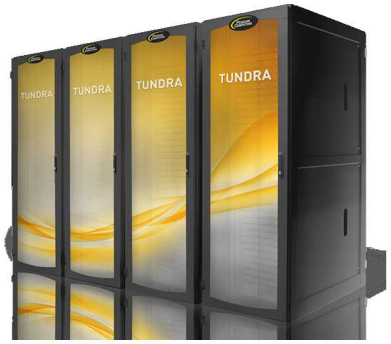
Joseph Kenny
Sandia National Laboratories
Livermore, CA

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

DOE Computing Ecosystem

Capability, e.g. Trinity
44PFlop/s

Institutional Capacity, e.g. CTS-1
0.8-2.8 PFlop/s



Institutional Non-traditional (CA)
Kahuna 108 Tflop/s, Carnac 310 TFlop/s

Individual/Departmental Resources
Anything and Everything

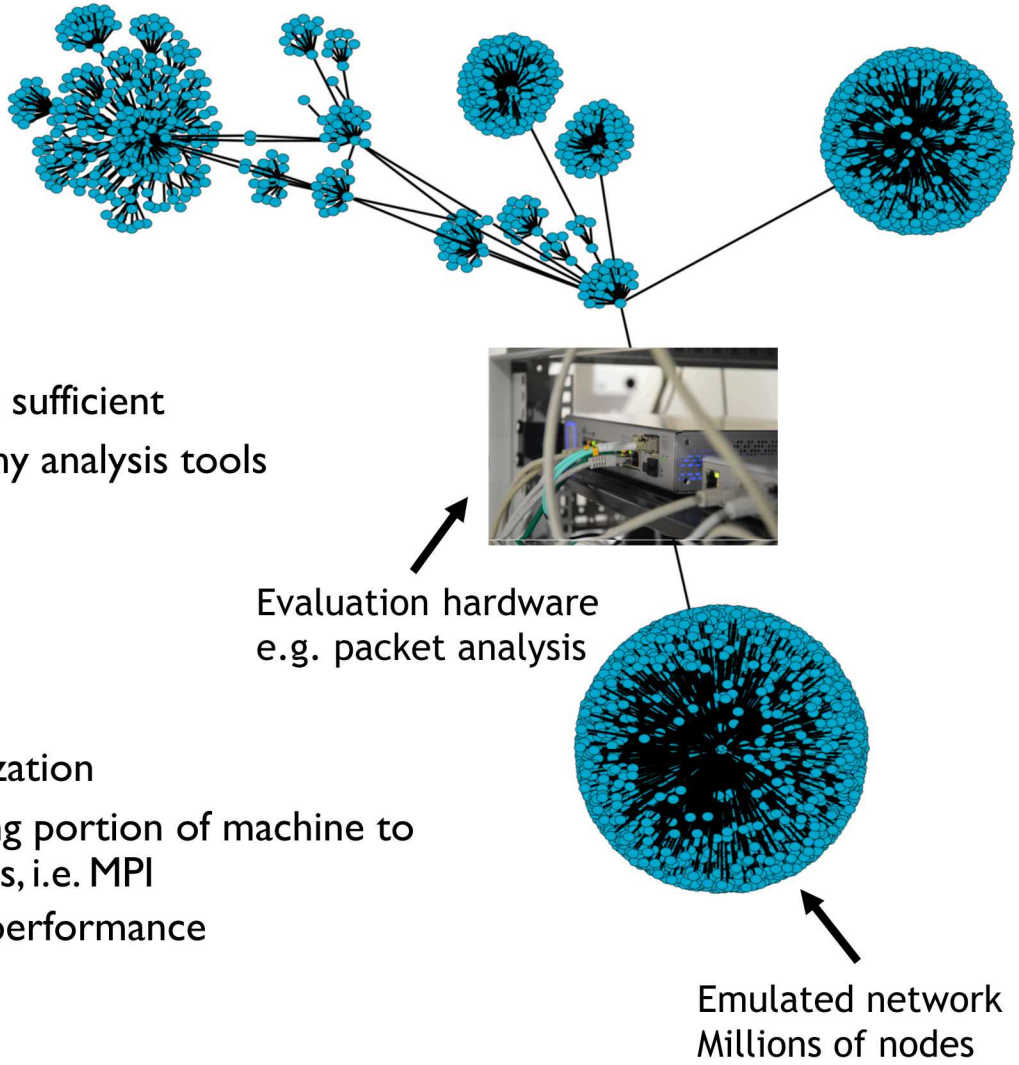
Non-traditional Systems (California Site)

Kahuna, Data Analytics (2015)
2 compute racks, 108 Tflop/s
NVMe Local Storage
Ceph, BeeGFS, Object Storage
Spark, R, Jupyterlab, Singularity
~150 users

Carnac, Network Emulation (2017)
6 compute racks, 310 Tflop/s
100Gb Ethernet
Bare metal provisioning
~50 users

(+ big GPU boxes and other random stuff…)

Network Emulation



100GbE Experiment Network

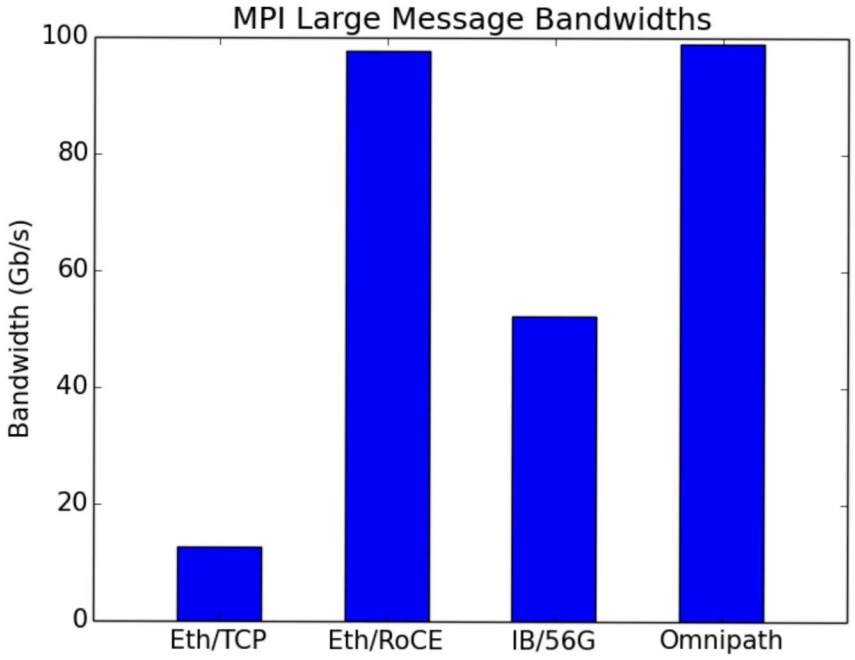
- Easy endpoint setup
- Many low bandwidth streams are sufficient
- Ethernet frames required for many analysis tools

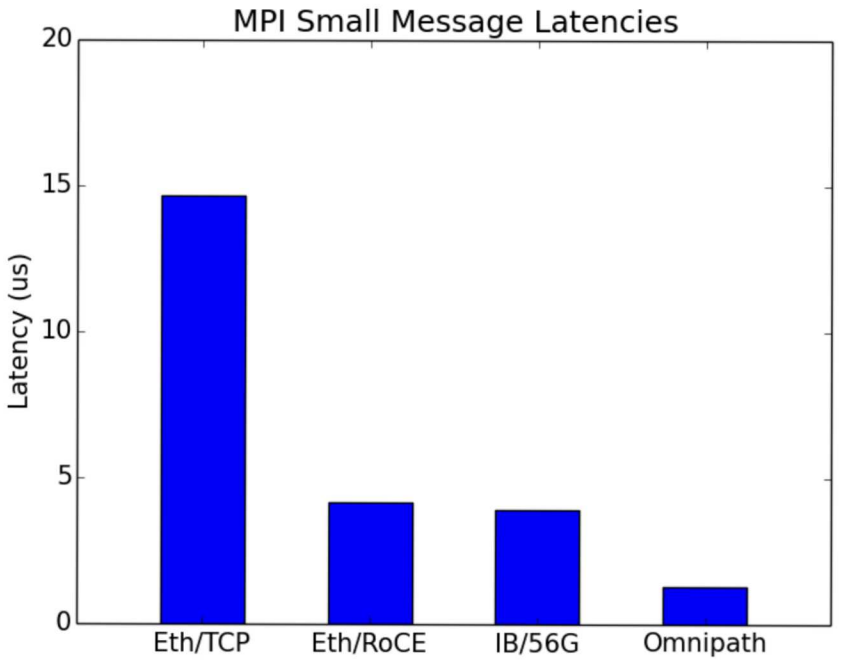
Small User Community

- Inconsistent machine utilization
- Would like option to swing portion of machine to traditional HPC workloads, i.e. MPI
  - ➔ need single stream performance

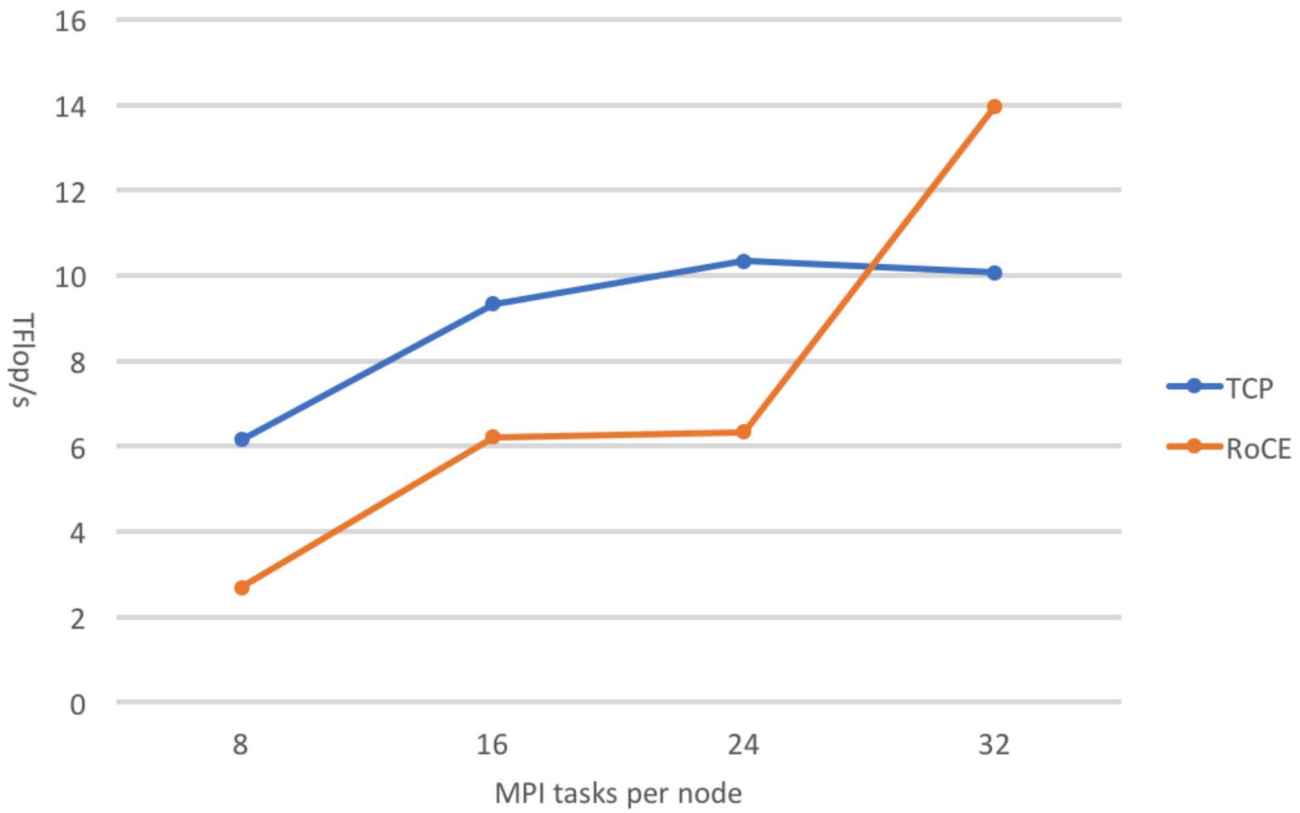
ROCE?

RoCE Single Stream MPI Performance





RoCE Linpack Performance



Conclusion

RoCE Experiences to date

- Documentation is poor
- At least some (v1) implementations handle heavy load poorly (or are difficult to configure well)
- Support for multivendor network hardware is painful

Work in Progress

- Better analysis of current architecture
- Evaluate single vendor end-to-end solution

More questions than answers at this point.
Here to “meet the experts” (commiserate).