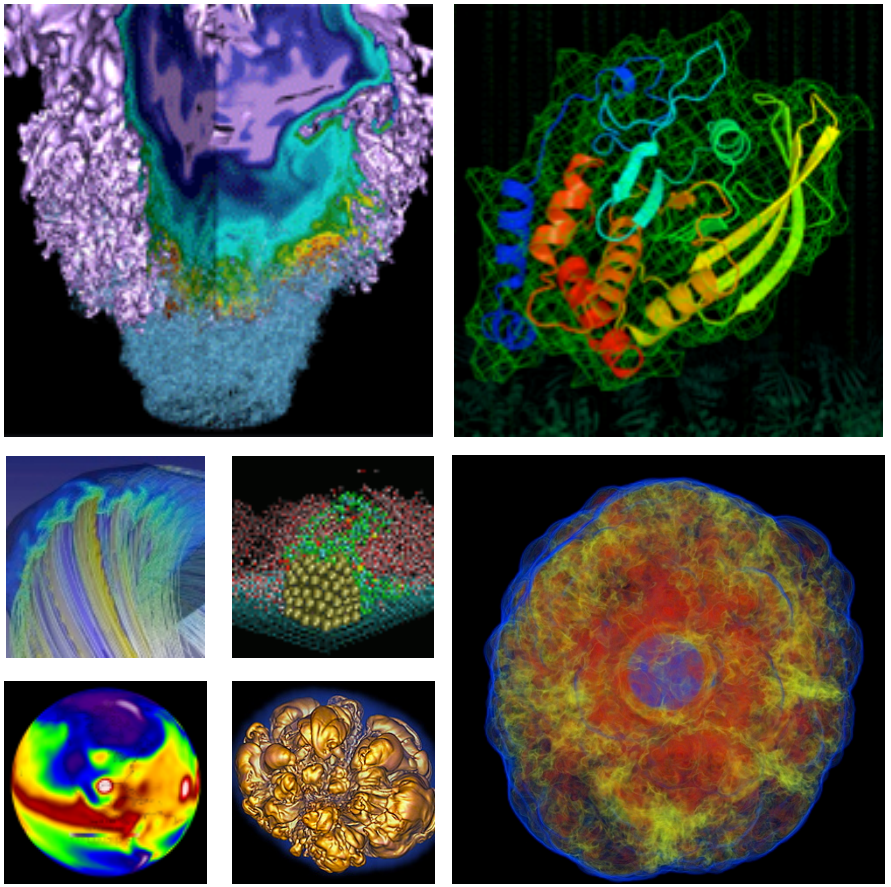
Characterization of the Cray Aries Network



Brian Austin
NERSC Advanced Technology Group

NUG 2014
February 6, 2014

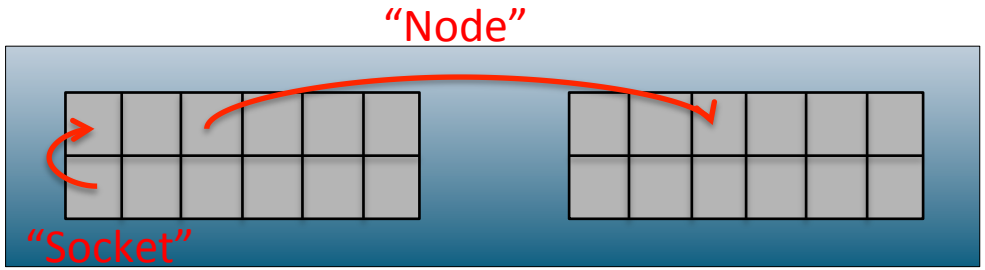
Edison at a Glance

- First Cray XC30
- Intel Ivy Bridge 12-core, 2.4GHz processors
  (upgraded from 10-core, 2.6GHz)
- Aries interconnect with Dragonfly topology
- Performs 2-4 x Hopper per node on real applications
- 3 Lustre scratch file systems configured as 1:1:2 for capacity and performance
- Access to NERSC's GPFS global file system via DVS
- 12 x 512GB login nodes
- Ambient cooled for extreme energy efficiency

On-node MPI point-to-point performance

- **On-node MPI performance is "speed of light" for interconnect performance.**
- **Measured using OSU MPI Benchmarks**
  - 8-byte latency
  - Bidirectional bandwidth
- **Two 12-core Intel Ivy Bridge processors per node.**

	Latency (us)	Bandwidth (GB/s)
Socket	0.3	
Node	0.7	



On-blade MPI point-to-point performance

- **A "blade" hosts four nodes and an Aries router.**
- **Nodes are connected to the Aries by PCIe-3 connections, each capable of 16 GB/s/dir.**

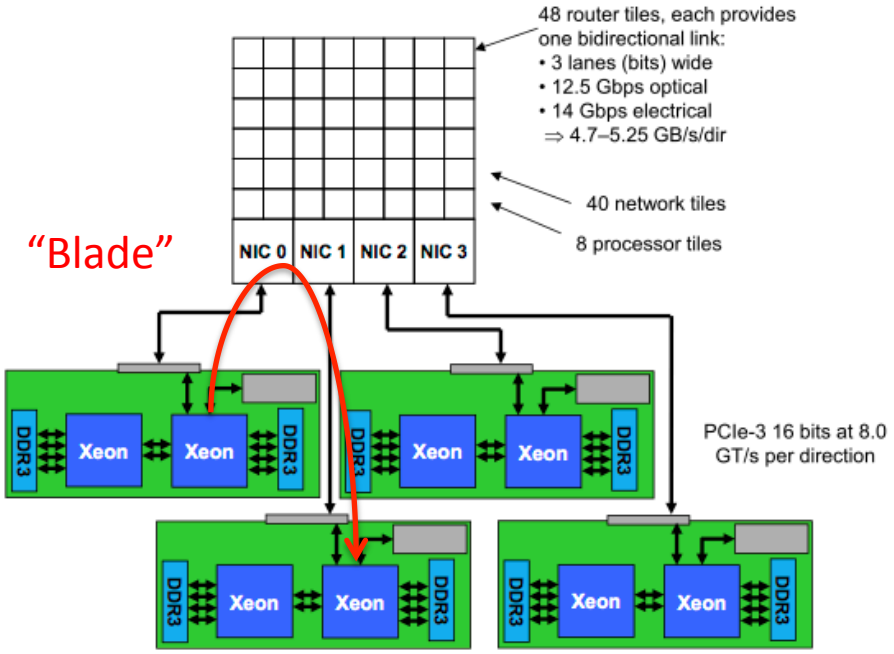
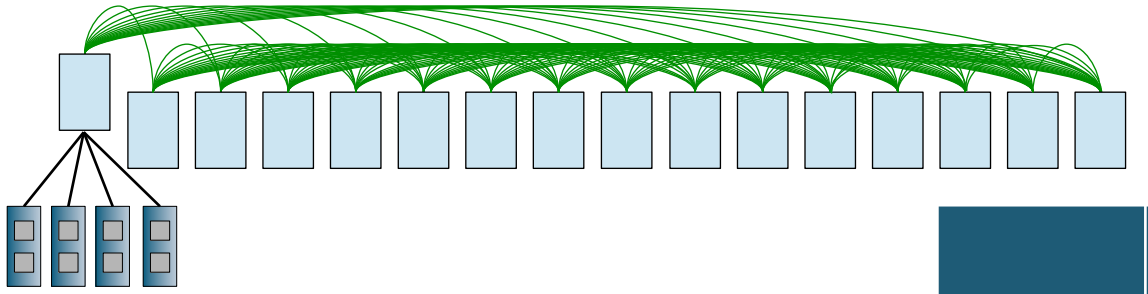


Image: Cray XC30 Network Guide

	Latency (us)	Bandwidth (GB/s)
Socket	0.3	
Node	0.7	
Blade	1.3	14.9
Rank-1	1.5	15.4

Rank-1 MPI point-to-point performance

- Sets of sixteen blades are packaged in a chassis.
- The rank-1 network (green) provides a direct connection between each pair of Aries within the chassis.
- Each rank-1 link provides 5.25 GB/s/dir.
- Measured BW exceeds link BW thanks to adaptive routing.



	Latency (us)	Bandwidth (GB/s)
Socket	0.3	
Node	0.7	
Blade	1.3	14.9
Rank-1	1.5	15.4

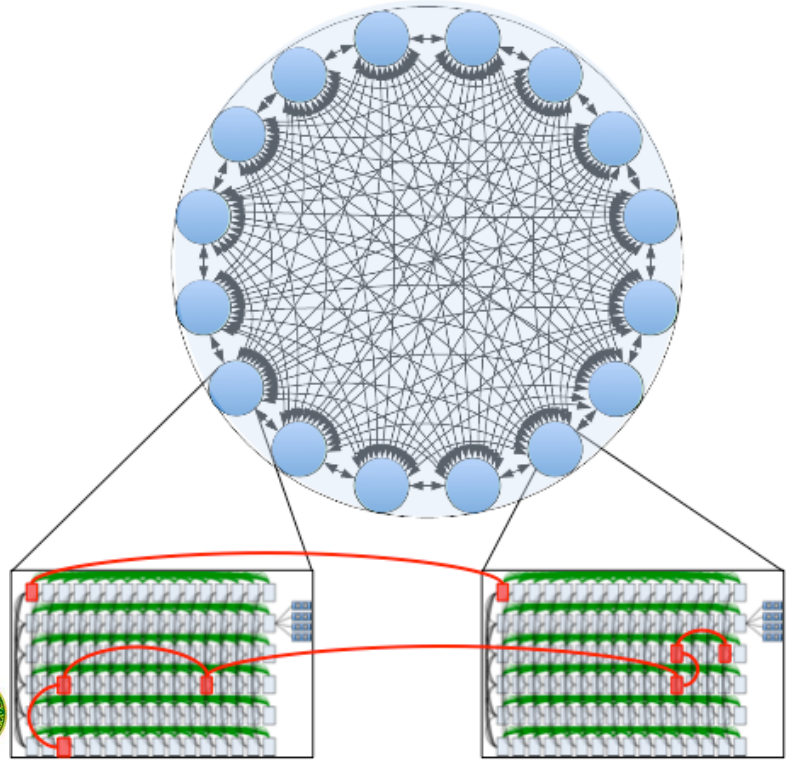
Rank-2 MPI point-to-point performance

- Sets of six chassis compose a 'group'.
- Within a group, the rank-2 network (black) connects each Aries to all of its peers in the other chassis.
- Each rank-2 connection provides 15.7 GB/s/dir.

	Latency (us)	Bandwidth (GB/s)
Socket	0.3	
Node	0.7	
Blade	1.3	14.9
Rank-1	1.5	15.4
Rank-2	1.5	15.4

Rank-3 MPI point-to-point performance

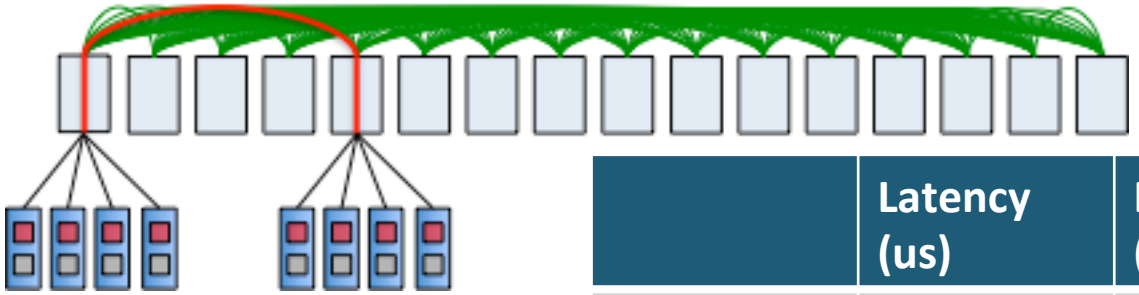
- Groups are connected by the "blue" rank-3 network.
- Rank-3 has all-to-all topology connected by optical links.
- Number of groups and inter-group bandwidth are configuration options. Edison has 14 groups (15 soon!) and 18.8 GB/s/dir per rank-3 connection.



	Latency (us)	Bandwidth (GB/s)
Socket	0.3	
Node	0.7	
Blade	1.3	14.9
Rank-1	1.5	15.4
Rank-2	1.5	15.4
Rank-3	2.2	15.3
Farthest	2.3	15.3

Point-to-point multi-bandwidth

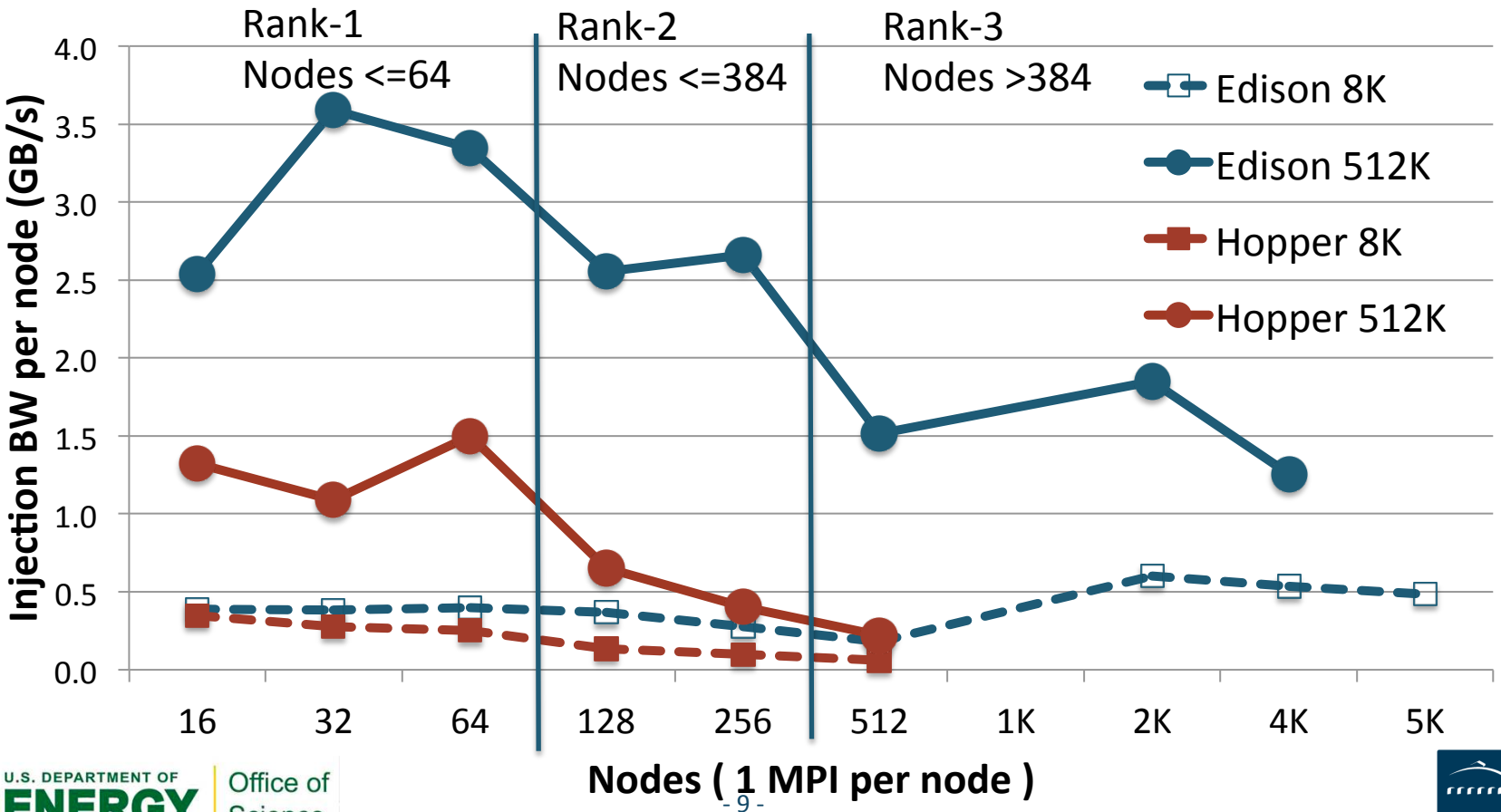
- **Point-to-point benchmark does not reflect bandwidth tapering in higher-rank networks.**
- **Limited by injection bandwidth at the NIC.**
- **To push the Aries network, we need multiple nodes to be active on each router (so Aries injection BW exceeds combined NIC injection BW).**



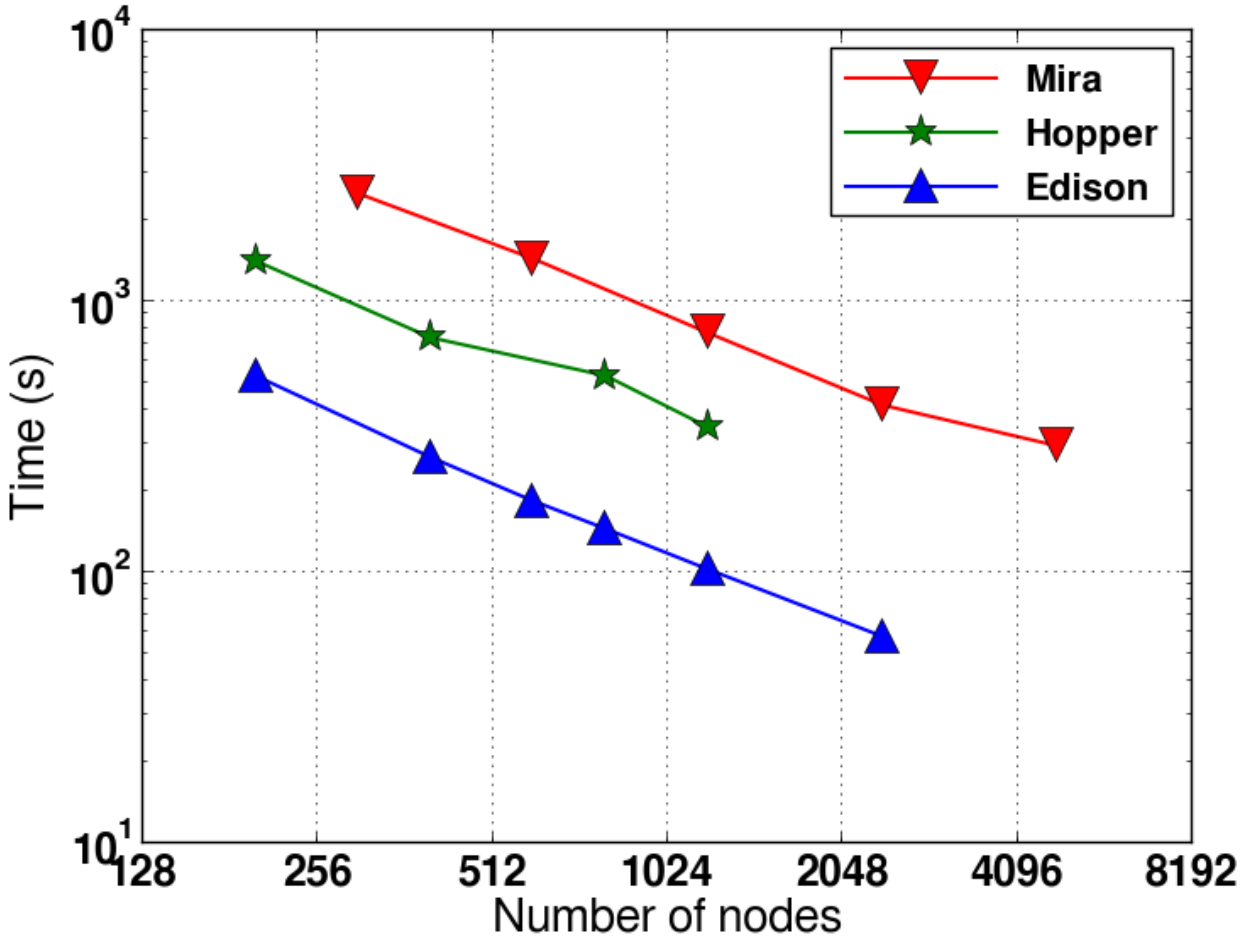
	Latency (us)	Bandwidth (GB/s)	Multi-BW (GB/s)
Rank-1	1.5	15.4	27.0
Rank-2	1.5	15.4	16.2
Rank-3	2.2	15.3	10.0
Farthest	2.3	15.3	5.1

Aries provides scalable global bandwidth.

- Within a group, MPI_Alltoall bandwidth is extremely high.
- Good alltoall bandwidth is sustained up to full system.

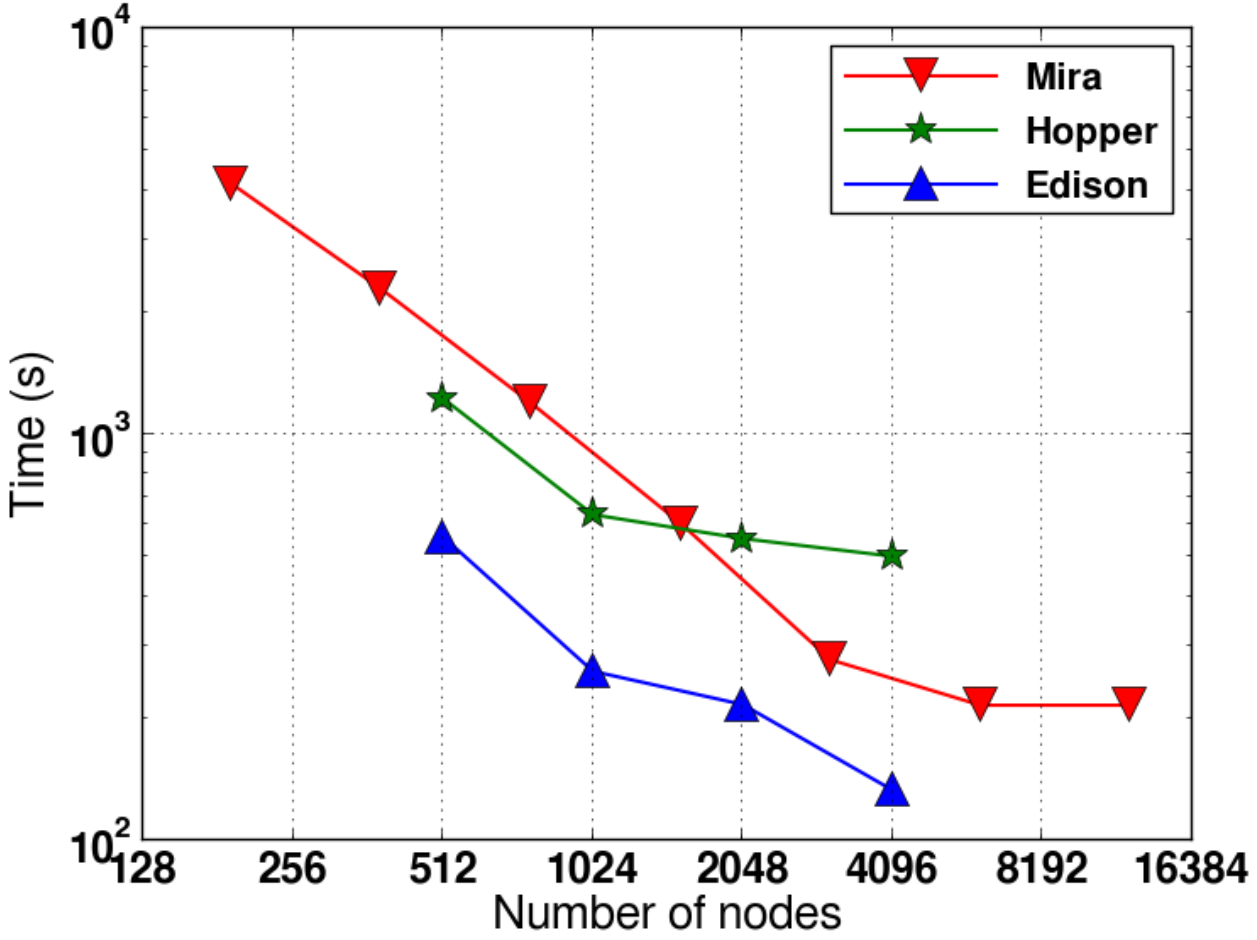


GTC



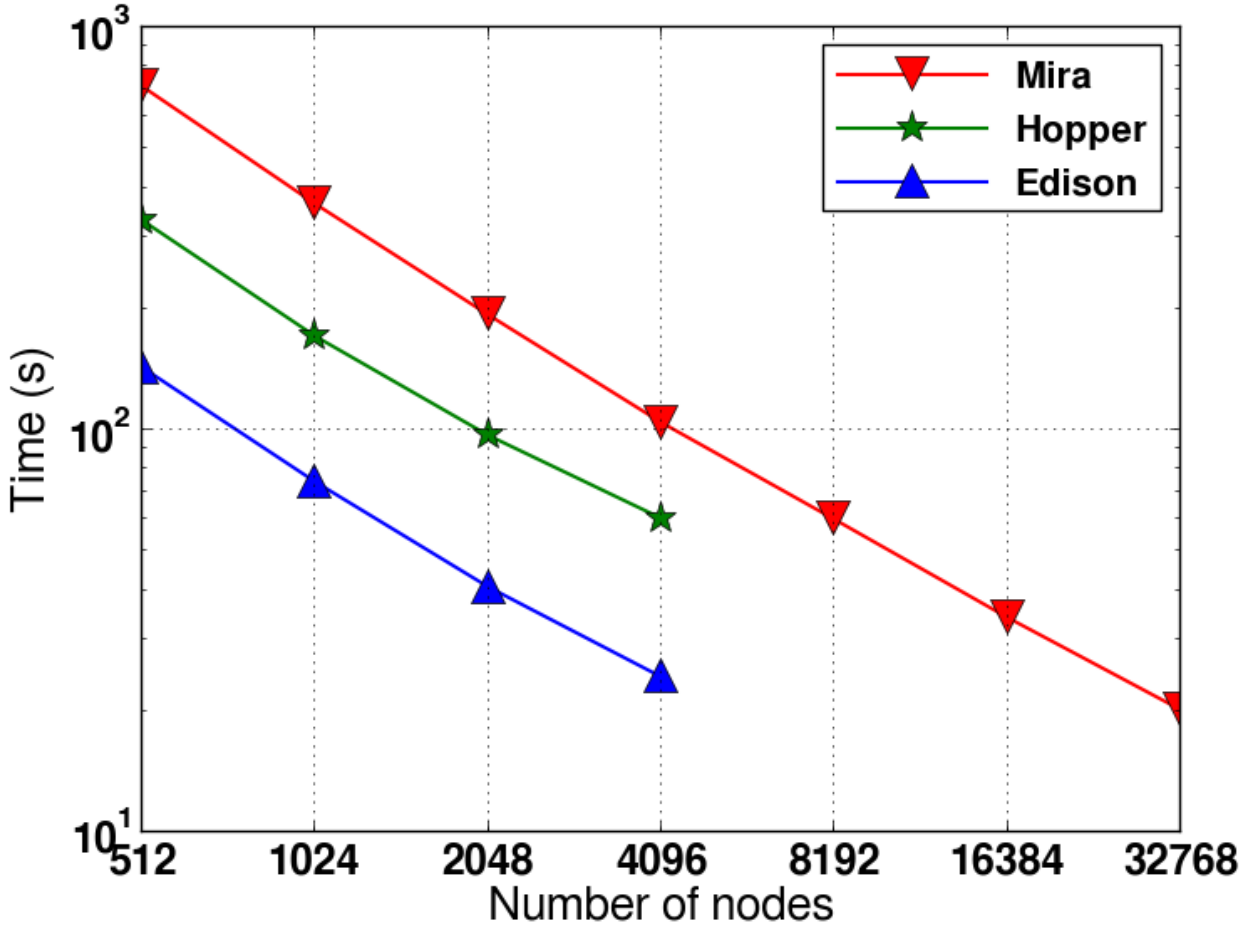
Edison performance: 2.0 – 2.6x Hopper, 7.0 - 7.8x Mira

MILC



Edison performance: 2.2 - 3.8x Hopper, 1.9 - 3.8x Mira

FLASH



Edison performance: 2.3 – 2.5x Hopper, 4.3 – 5.0x Mira

Machine comparison summary

- **At a given node count, the best application runtime was always achieved on Edison**
- **Need ~4x Mira nodes to improve upon an Edison time**
  - The Mira nodes are more power efficient, but you must work harder to identify more parallelism to run well on Blue Gene platforms
- **Edison's Aries interconnect helps application scalability**
  - In our NERSC-8 applications, we saw several examples of loss of scalability at very-large scale on Hopper which did not happen on Edison

Thank you.