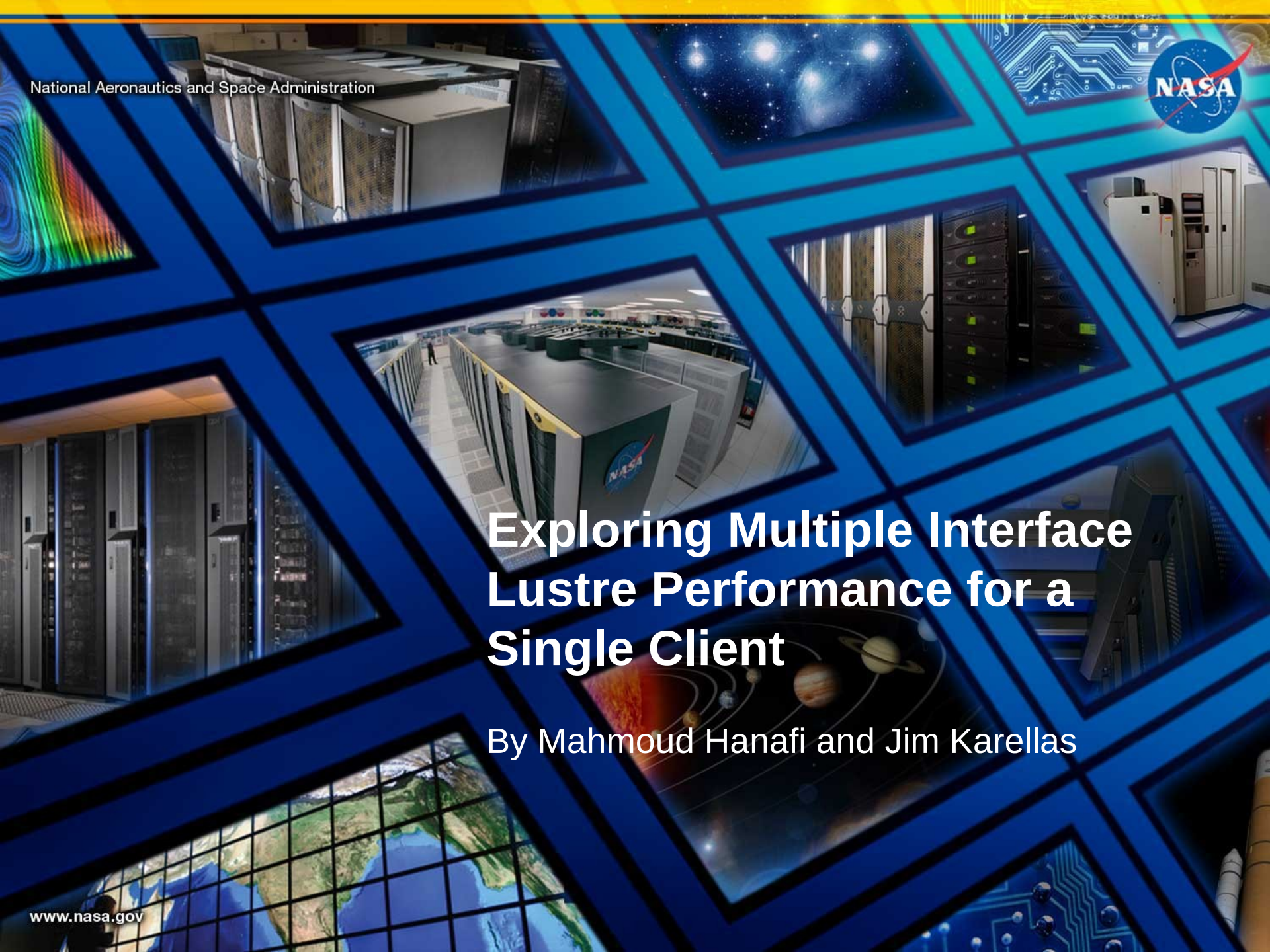

National Aeronautics and Space Administration
NASA
Exploring Multiple Interface Lustre Performance for a Single Client
By Mahmoud Hanafi and Jim Karellas
www.nasa.gov

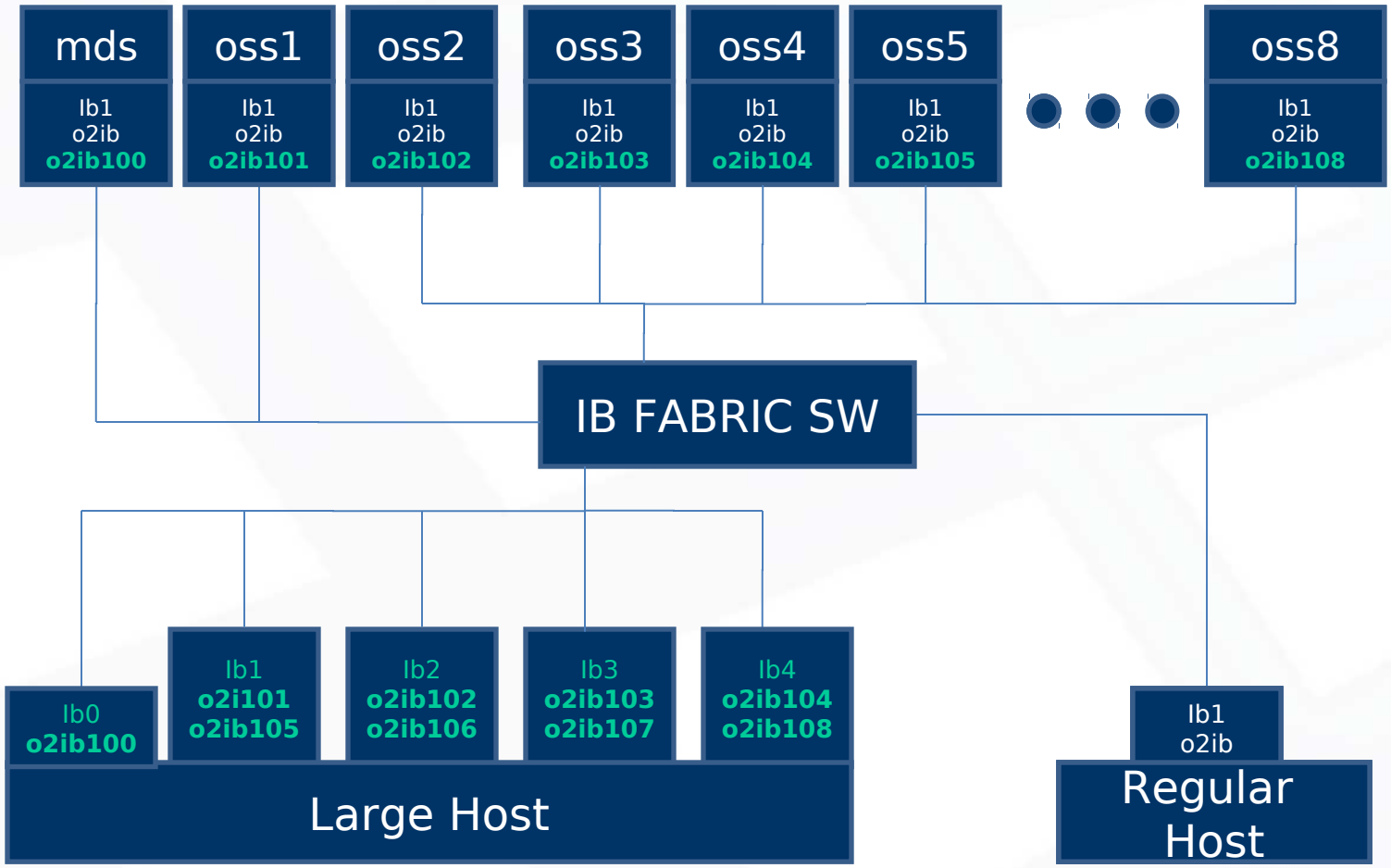
Motivation

- Multiple File Systems
- Multiple IB Fabrics
- Large Node Single Clients
- NASA UV 2000
  - Processors: Intel Xeon E5-4650L
  - Cores: 1024
  - Memory: 4TB
- Old Altix based SSI system: 8 SDR cards=3GB/s

Requirements

- Control IB connection for each OSS (multiple NIDS)
- Ability to add as many interfaces as needed per-client
- No need for special IP range addressing

NID Layout (Generic System)



Endeavour2 NIDs (rails 2-6)

RAIL2

ib1	ib2
o2ib100(ib1)	
o2ib101(ib1)	o2ib102(ib2)
o2ib103(ib1)	o2ib104(ib2)
o2ib105(ib1)	o2ib106(ib2)
o2ib107(ib1)	o2ib108(ib2)
o2ib109(ib1)	o2ib110(ib2)
o2ib111(ib1)	o2ib112(ib2)

RAIL3

ib1	ib2	ib3
o2ib100(ib1)		
o2ib101(ib1)	o2ib102(ib2)	o2ib103(ib3)
o2ib104(ib1)	o2ib105(ib2)	o2ib106(ib3)
o2ib107(ib1)	o2ib108(ib2)	o2ib109(ib3)
o2ib110(ib1)	o2ib111(ib2)	o2ib112(ib3)

RAIL4

ib1	ib2	ib3	ib4
o2ib100(ib1)			
o2ib101(ib1)	o2ib102(ib2)	o2ib103(ib3)	o2ib104(ib4)
o2ib105(ib1)	o2ib106(ib2)	o2ib107(ib3)	o2ib108(ib4)
o2ib109(ib1)	o2ib110(ib2)	o2ib111(ib3)	o2ib112(ib4)

RAIL5

ib1	ib2	ib3	ib4	ib5
o2ib100(ib1)				
o2ib101(ib1)	o2ib102(ib2)	o2ib103(ib3)	o2ib104(ib4)	o2ib105(ib5)
o2ib106(ib1)	o2ib107(ib2)	o2ib108(ib3)	o2ib109(ib4)	o2ib110(ib5)
	o2ib111(ib2)	o2ib112(ib3)		

RAIL6

ib1	ib2	ib3	ib4	ib5	ib6
o2ib100(ib1)					
o2ib101(ib1)	o2ib102(ib2)	o2ib103(ib3)	o2ib104(ib4)	o2ib105(ib5)	o2ib106(ib6)
o2ib107(ib1)	o2ib108(ib2)	o2ib109(ib3)	o2ib110(ib4)	o2ib111(ib5)	o2ib112(ib6)

Endeavour2 modprobe.d/lustre (Example)

```
# 1 PORT

options lnet networks=o2ib(ib1)

-------------------------------------------------------------------------

# 4 PORTS
options lnet
   networks=o2ib100(ib1),o2ib101(ib1),o2ib102(ib2),o2ib105(ib
   1),o2ib106(ib2),o2ib109(ib1),o2ib110(ib2),o2ib103(ib3),o2ib1
   04(ib4),o2ib107(ib3),o2ib108(ib4),o2ib111(ib3),o2ib112(ib4)
```

BENCHMARKS

- IOR options
  - POSIX
  - Each task read/writes 1 striped file per OST
  - 1MB Block Size
  - 1G file
  - Direct I/O

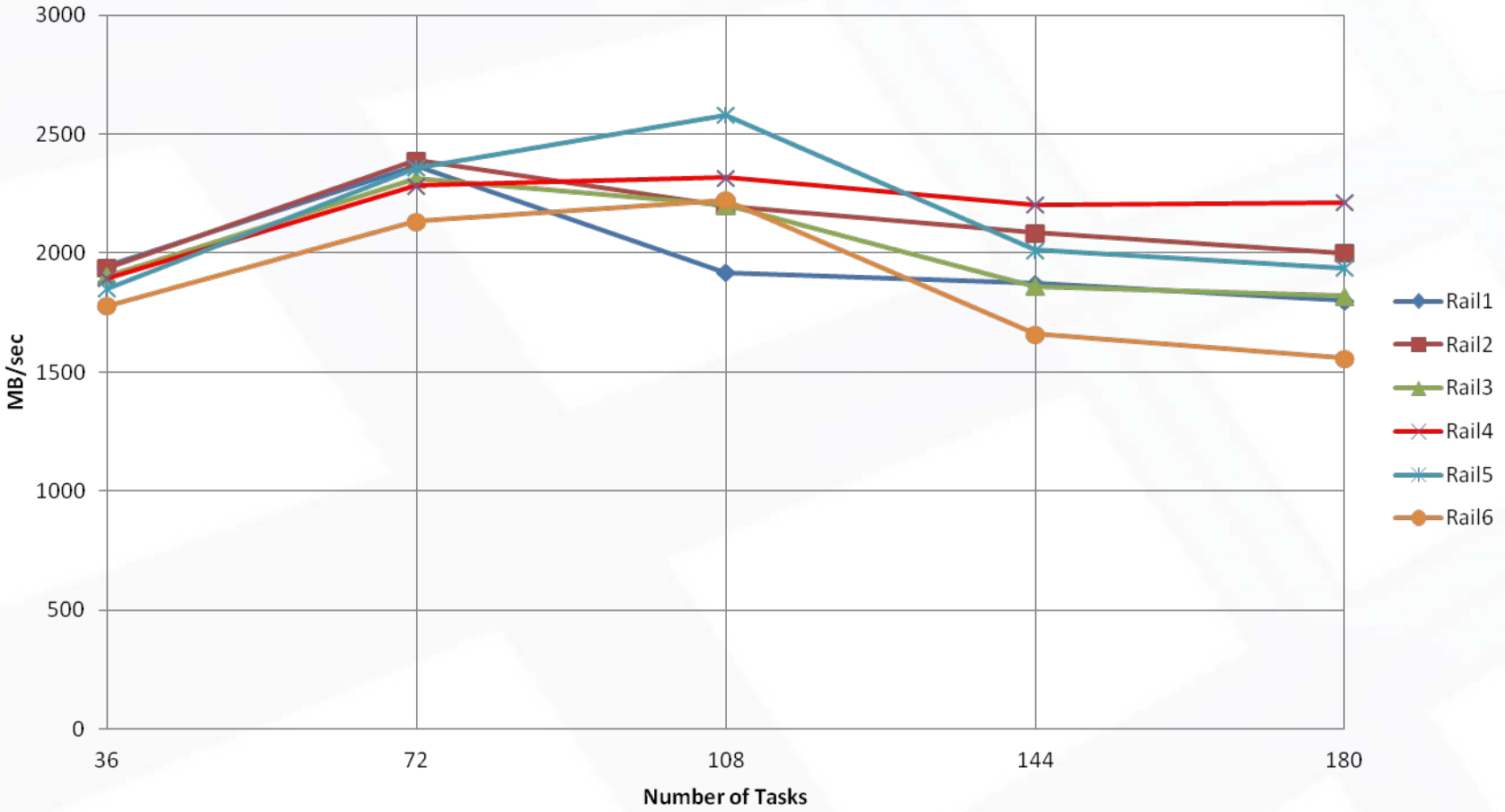
- File system Backend Netapp e5400
  - Able to do > 7GB/sec

Software Versions

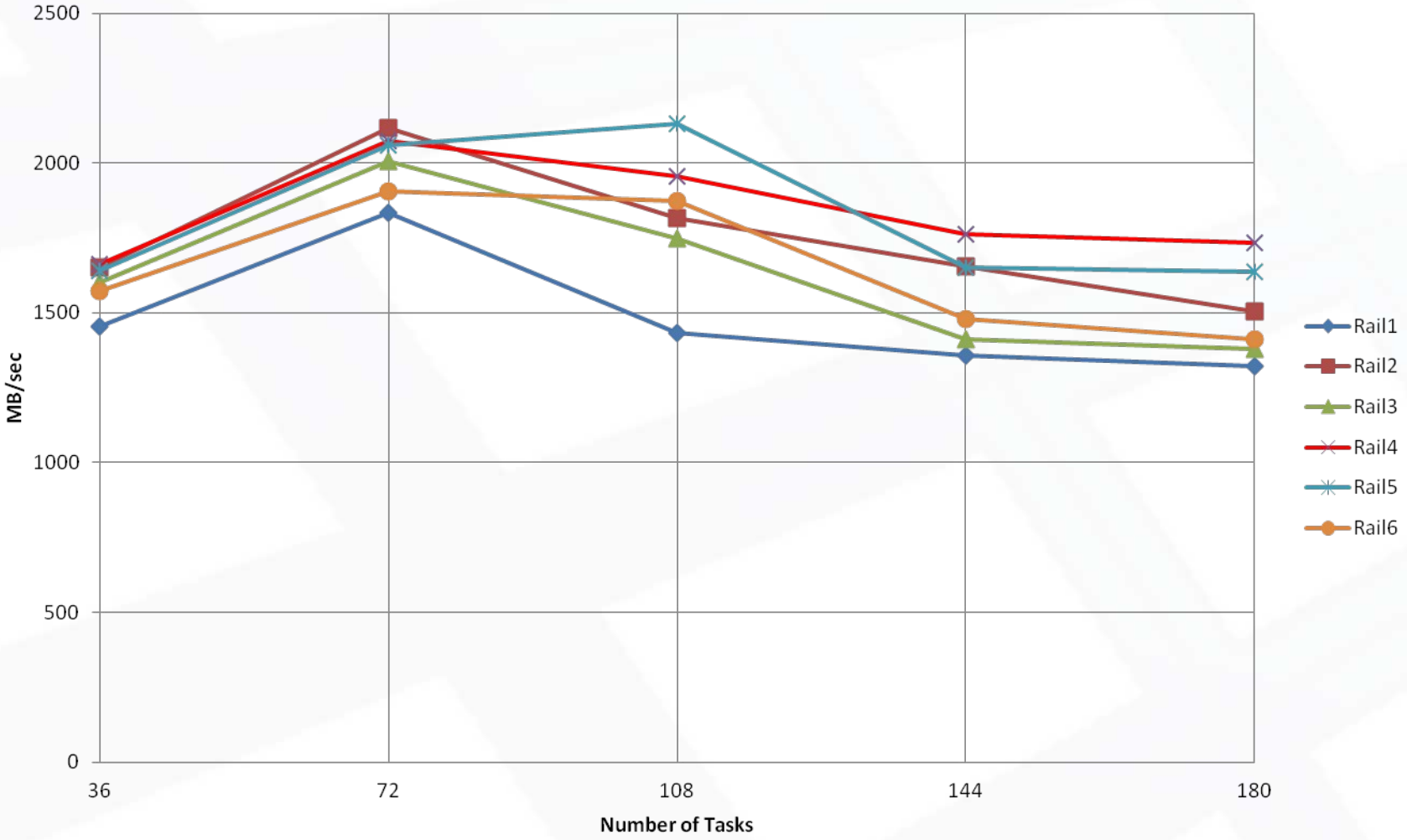
- Lustre Servers
  - Centos6.3
  - Kernel: 2.6.32-279.19.1
  - Lustre 2.1.4 (with additional patches)
  - OFED 1.5.4

- Lustre Client
  - Sles11SP2
  - Kernel: 3.0.51-0.7.9.1
  - Lustre 2.3.0-2
  - OFED 1.5.4

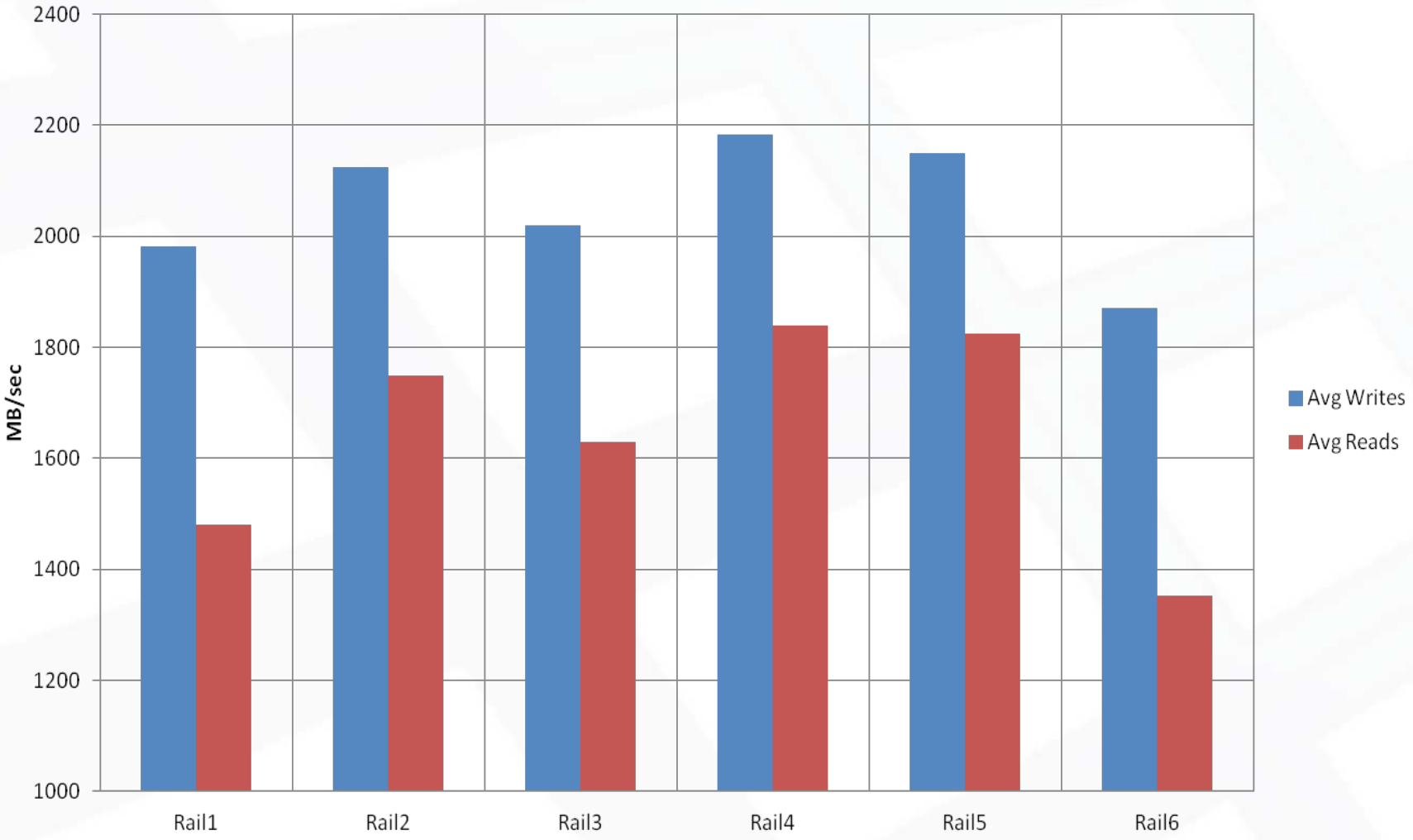
Sequential Writes



Sequential Reads



Average Over Tasks
(Tasks 36,72,108,144 and 180)



Conclusions/Future Work

- Scaling results were disappointing.
- Rail4 produced the best results.
- Investigate LNET tuning.
- Review Lustre client code.
- HCA placement in NUMA/SSI system is relevant.
- Connection location in IB fabric relevant, needs analysis.