Disk Storage performance & high-speed interconnects

Andreas Hirstius:

Test of a storage solution based on external USB 2.0 or FireWire disks

Andras Horvath:

Fast local interconnects

Péter Kelemen:

Disk server performace improvements

Mass Storage

Test of a storage solution based on external USB 2.0 or FireWire disks

Why do we look at it?

- 25PB of tape with 4GB/s   28 MCHF
- 28MCHF   14PB of mirrored disk (~100GB/s)
  (both from re-costing LCG phase 1+2 exercise predictions)
- "low end" storage w/o mirror: 28PB
- ~40PB when reducing cost for servers
- Power budget
  - ~100kW for tape infrastructure
  - ~100kW for 25000 disks (powersave <4W)

Why do we look at it? cont.

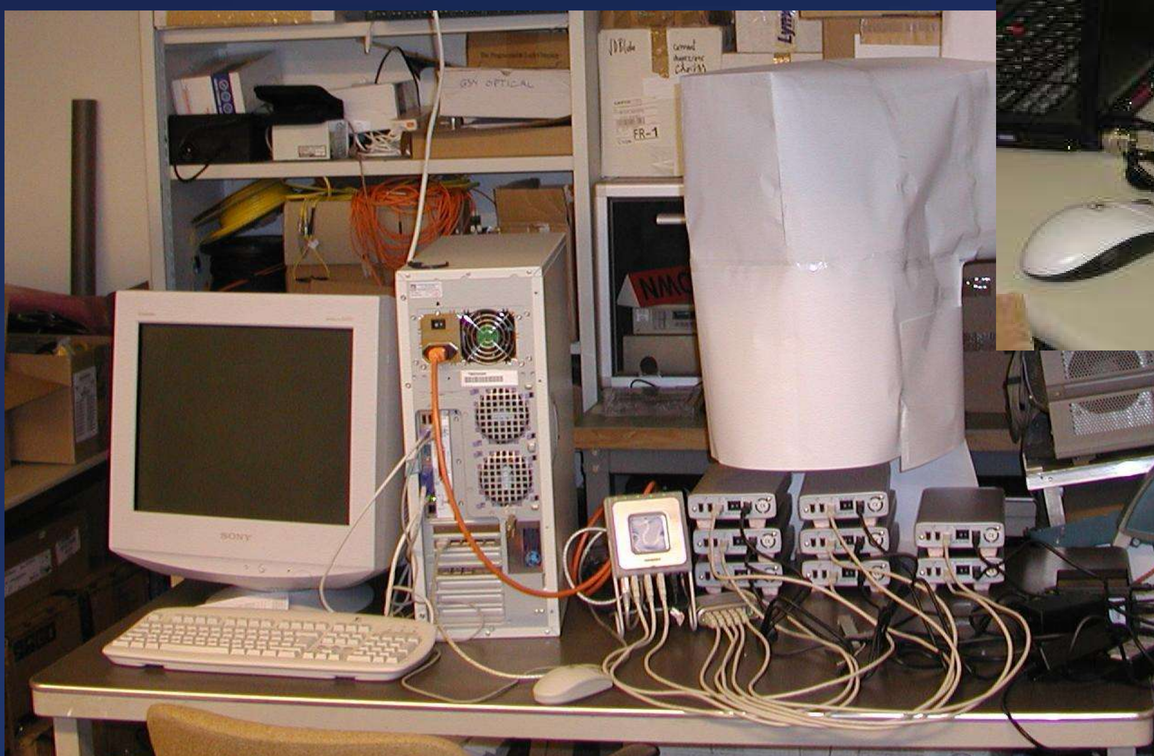
- Possible application
  - AddOn/Replacement for tapes for certain applications
    - Large volume – long lived data with rare access
    - Possibility to guarantee bandwidth to “tape”
- What do we want to check:
  - Is the throughput reasonable?
  - What about stability?
  - What happens when a problem occurs (disk dies)?

Details of the test setup

	USB 2.0	FireWire
Theor. throughput	480Mbit/s	400/800Mbit/s
Max. # of devices	127	63
Preferred structure	Tree	Daisy chain(tree possible)

- Dual Xeon and Laptop (IBM R40)
- USB 2.0 and FireWire controller (PCI or on-board)
- Different USB/FW hubs
- “old” IBM HDD in external case with Genesys chipset
- Maxtor OneTouch 250GB

The Test Setup



Power consumption

	1 stream	2 streams	4+ streams
Seq. Read	<15W	<15W	<15W
Seq. Write	<14W	<14W	<14W
Random read	<15W	~15W	**~16W**
Random write	<14W	<14W	<14W
Mkfs		~15W	
Startup		**<25W**	
Idle		**11-13W**	
Powersave		**<4W**	

max. power consumption for a single disk

Results with USB 2.0

- Single disk Transfer rate
  - Laptop: 24MB/s read; 26MB/s write
  - Dual Xeon: 27MB/s read; 23MB/s write
- Max. transfer rate (multiple disks)
  - Laptop: 43MB/s read and write
  - Dual Xeon: 28MB/s read and write
- The transfer rates are limited by the host controller only!

Maintainability and Stability

Main Problem: A system with ~100 external disks has to be stable when disk state is changing (adding, removing, etc.).

- Problematic disks automatically “disappeared”
  - Killing of application sometimes necessary
- Discovery of newly connected devices works fine
- Removal of unmounted disk w/o problem
- Stable long term behaviour
- Mount-by-label necessary

Problems seen with USB

- Serious problem on SMP machines
  - Any access to even a single disk causes kernel Oops
  - Forwarded to maintainer of the code
  - Not understood
- Genesys chipset in external case is buggy
  - workaround available, but performance degraded
  - "positive" effect: high failure rate was perfect for maintainability/stability tests

Results with FireWire

- No measurements with Genesys chipset possible
- Simple tests with the Maxtor OneTouch
  - Read and write transfer rates < 20MB/s
  - Standard procedures work fine
    - Removal of unmounted disk
    - Discovery of newly connected disk
- ALEPH used daisy-chained FireWire disks (>10)for data export

Outlook

- Consumer market has a clear trend to external disks

  (storage for Video, Music; user-friendly backup; trend towards Laptops)
- If this is seen as a viable option:
  - Large scale test system necessary
  - Problems expected in the kernel
    - Nobody has ever connected ~100 external disks to a box!
  - Development of the necessary software
    - Stager module
    - Disk management/monitoring, etc.

Fast local interconnects — Practical experience with Infiniband

Andras.Horvath@cern.ch

- Motivations
- Protocol stacks overview
- RFIO test results
- SDP (socket) first test results
- Conclusion so far
- Next steps

Motivations

- I/O intensive application
- More CPU power per node
- Faster storage

→ Need for faster interconnect

However...

- Current interconnects “expensive”
- Software-only protocols don't scale
- Need for resilience
- Need for open standards

Technologies overview

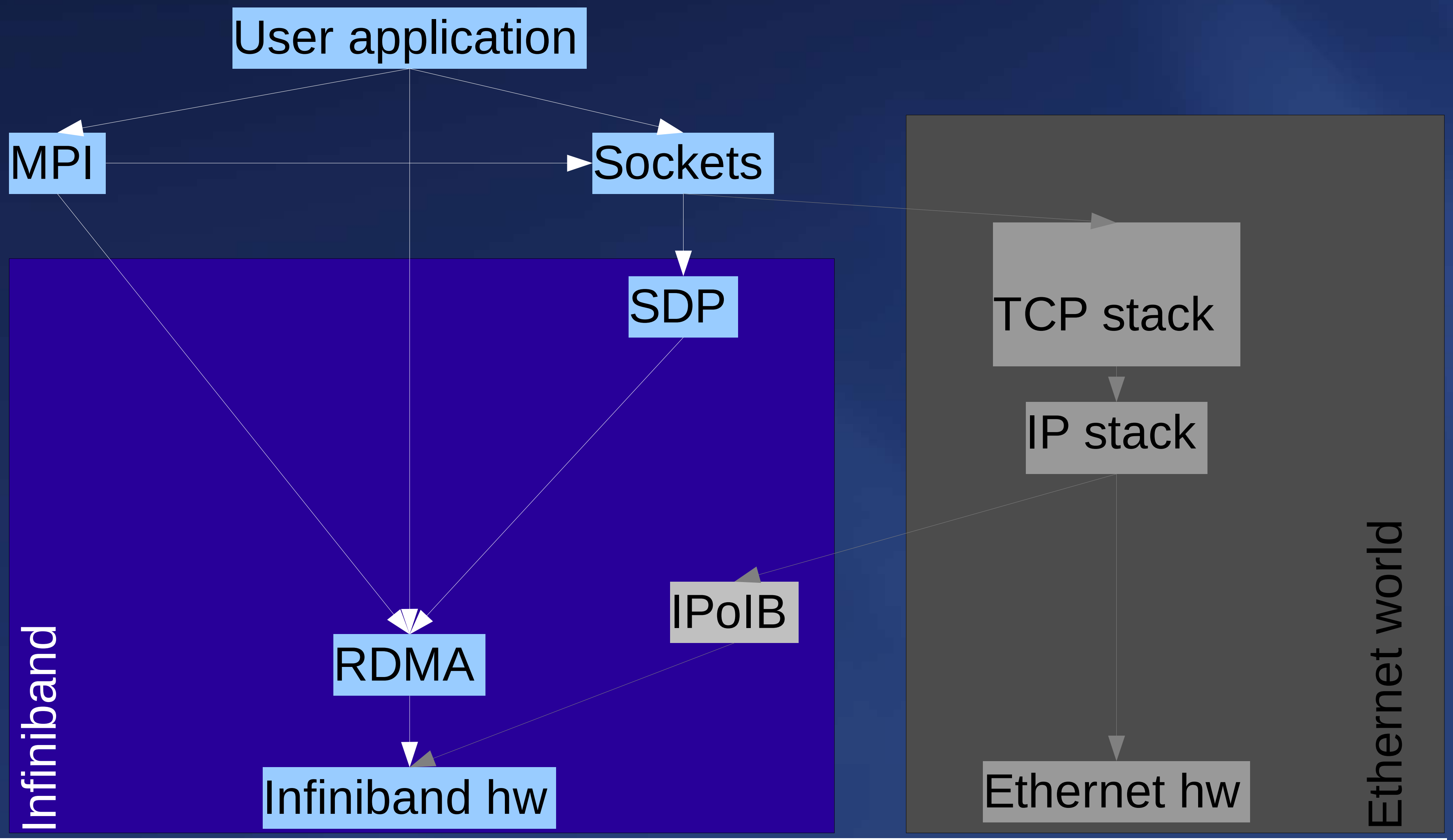
	Max bw, Gbit/s	Price/port, $	Proprietary	RDMA capable
Infiniband	10 and 30	1400	No	Yes
10Gigabit Ethernet	10	8000	No	No
Quadrics Elan4	10	2500	Yes	Yes
Fibre Channel	2	3000	No	No
Myrinet	2.5	1200	Yes	Yes

Remote Direct Memory Access (RDMA): technology to transfer data directly to/from remote address space *by the network hardware, not the CPU*

	TCP/IP and Ethernet	**Infiniband**
Data integrity preserved by...	host CPU (+ offload)	hardware
Unit of transfer	1500 byte	up to 4MB*
Data path redundancy by...	STP or routing	Fabric management
Switchover time	~1min	~1ms

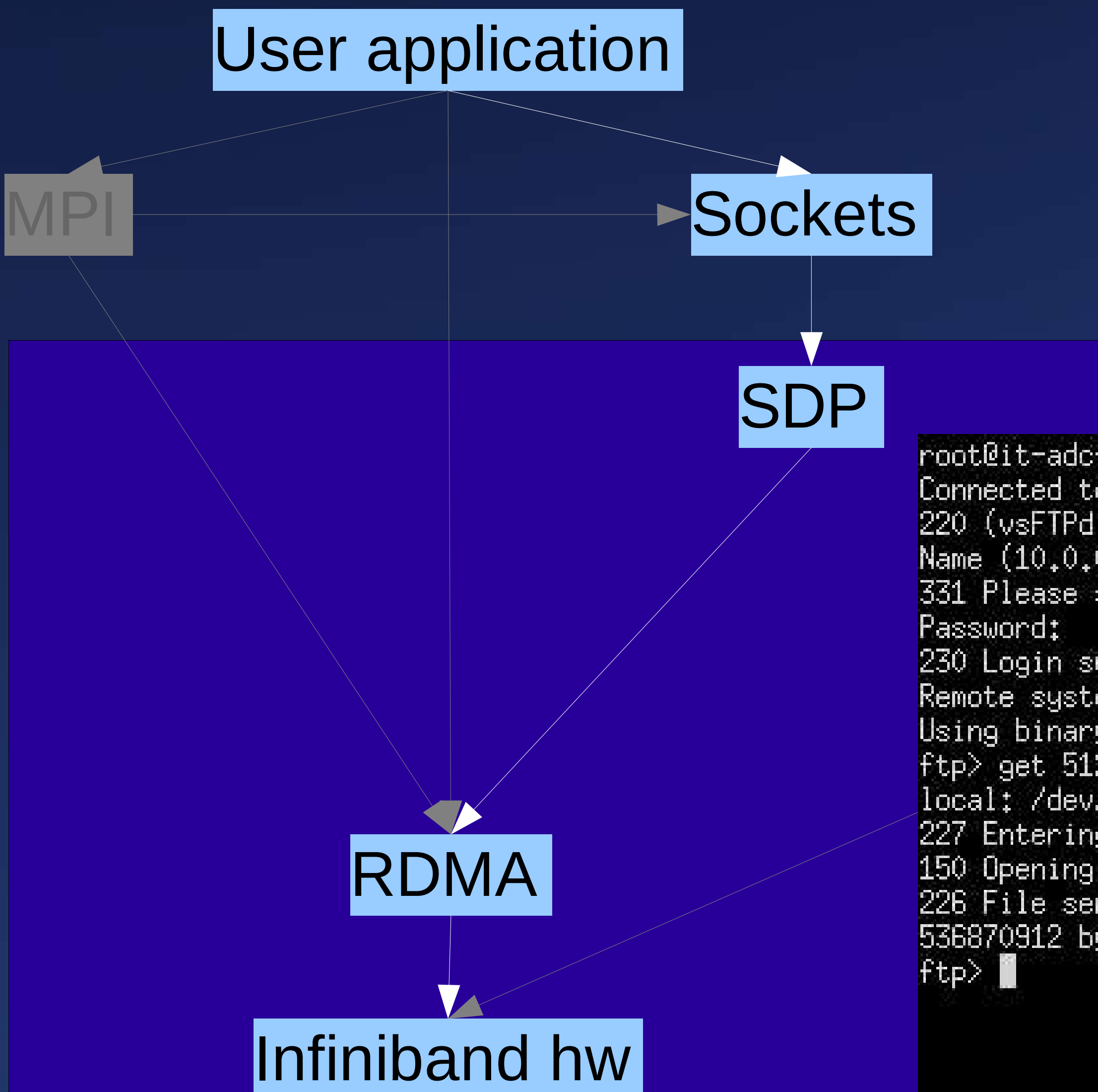
*: sockets via SDP: 2044(4092) byte, RDMA and MPI: 1byte - 4MB

Running applications

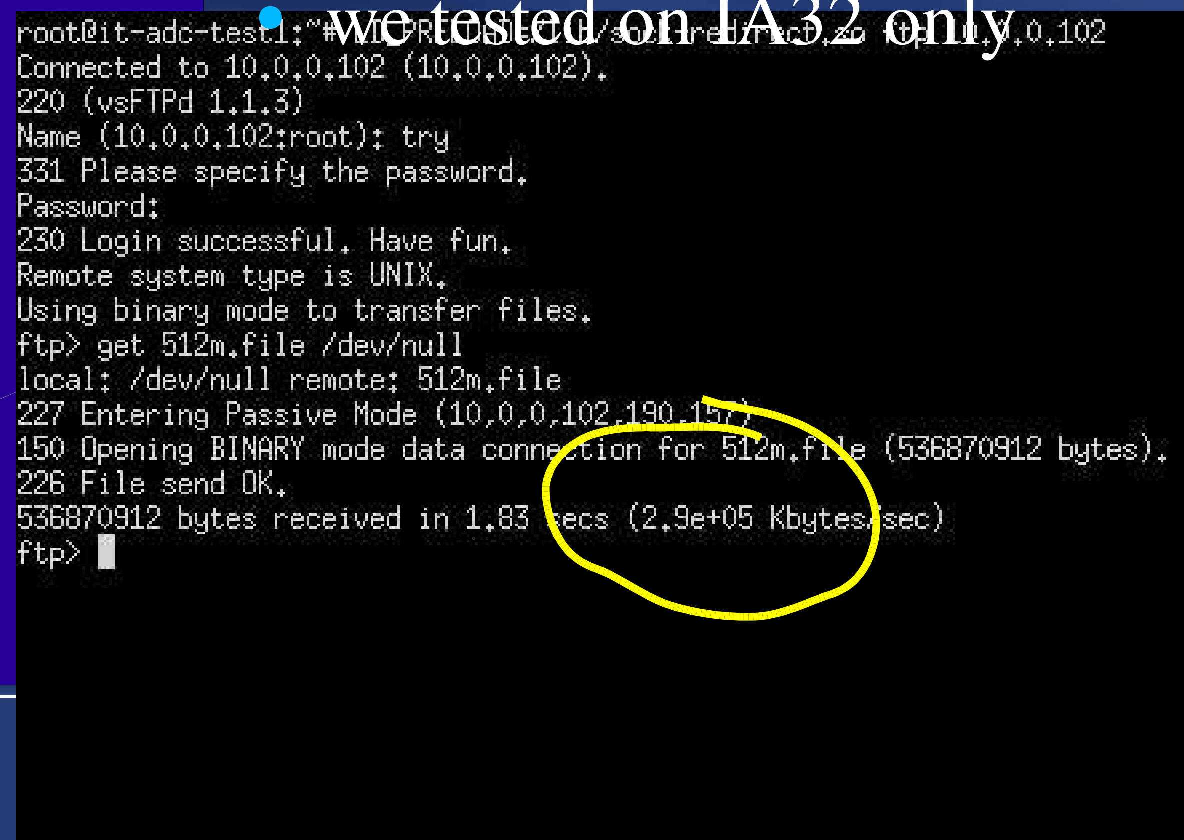


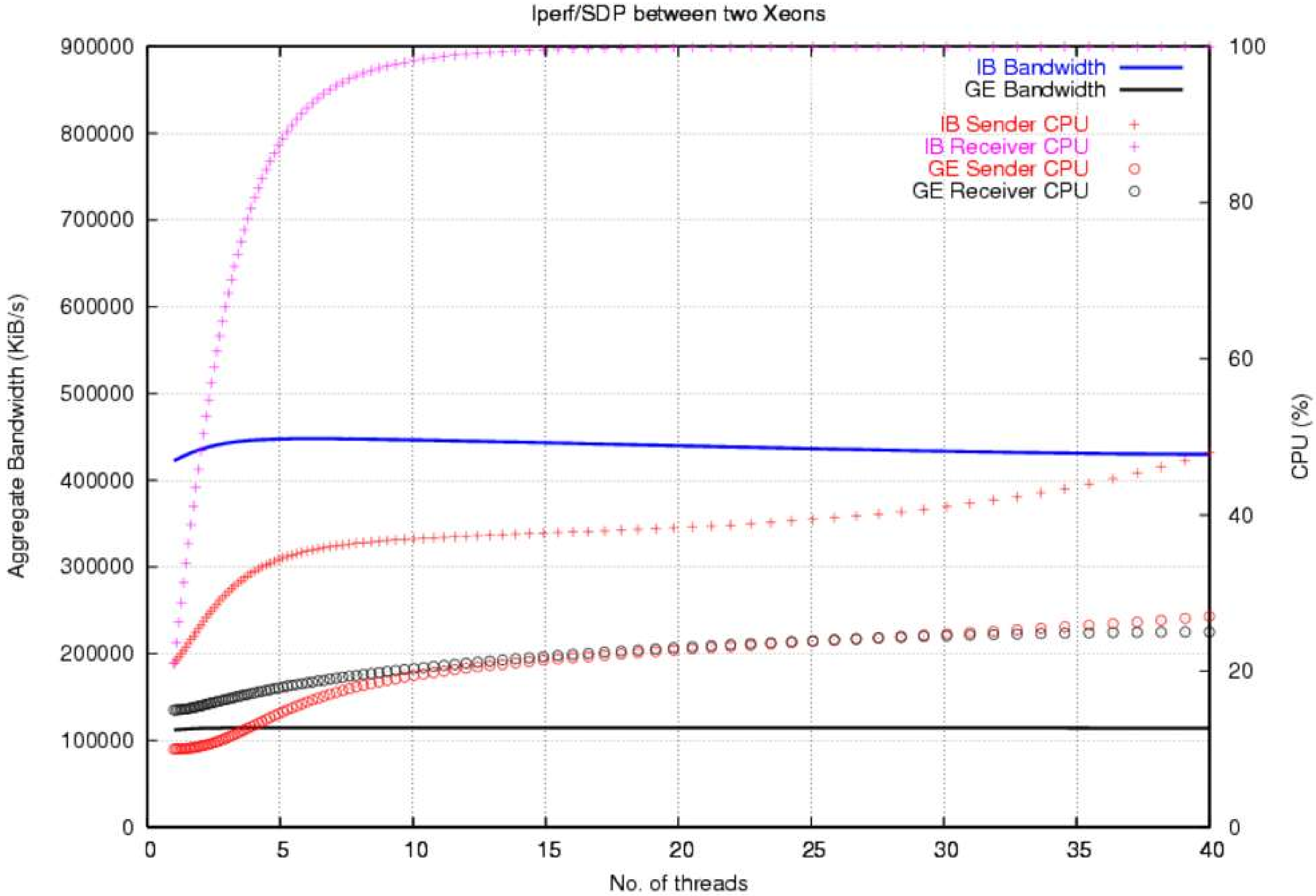
Andras Horvath

Sockets on IB: SDP



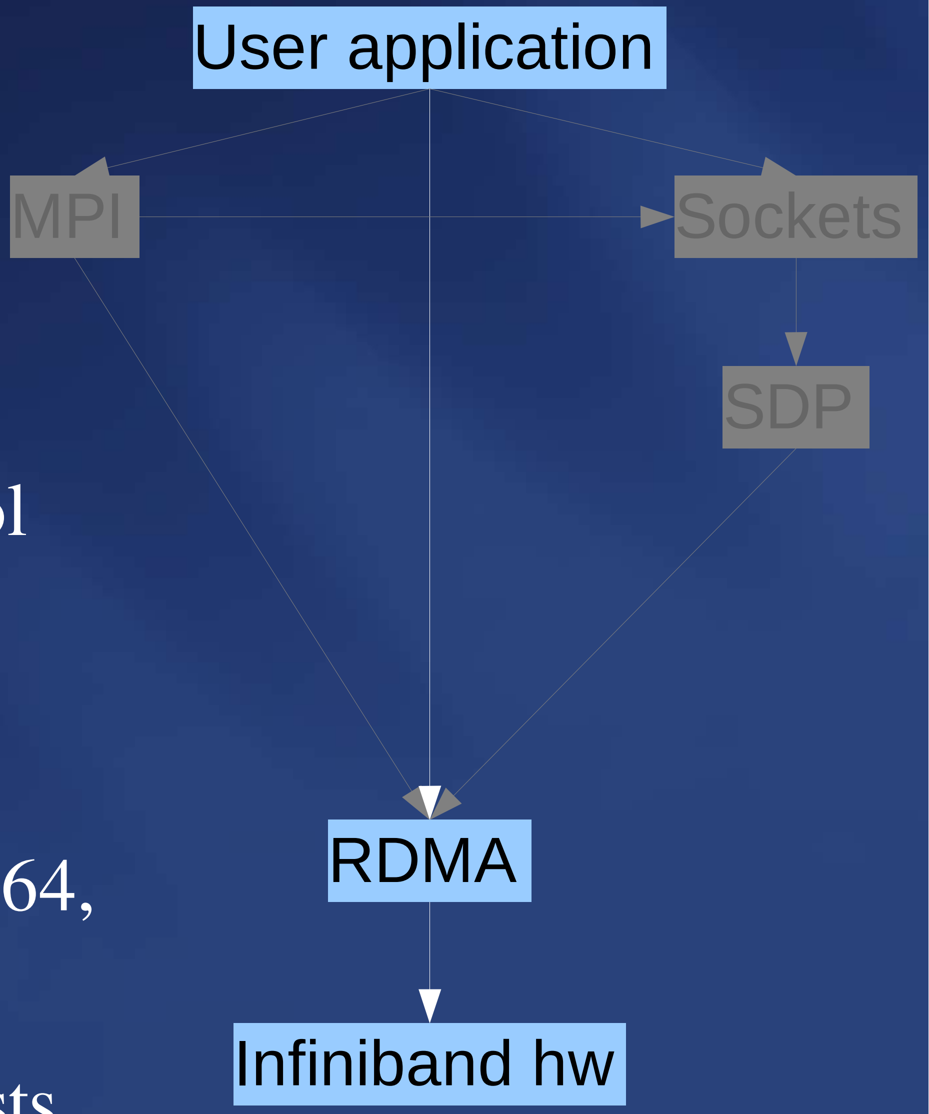
- Addressing via IP
- Use new BSD address family
- ... or LD_PRELOAD a librar
- we tested on IA32 only

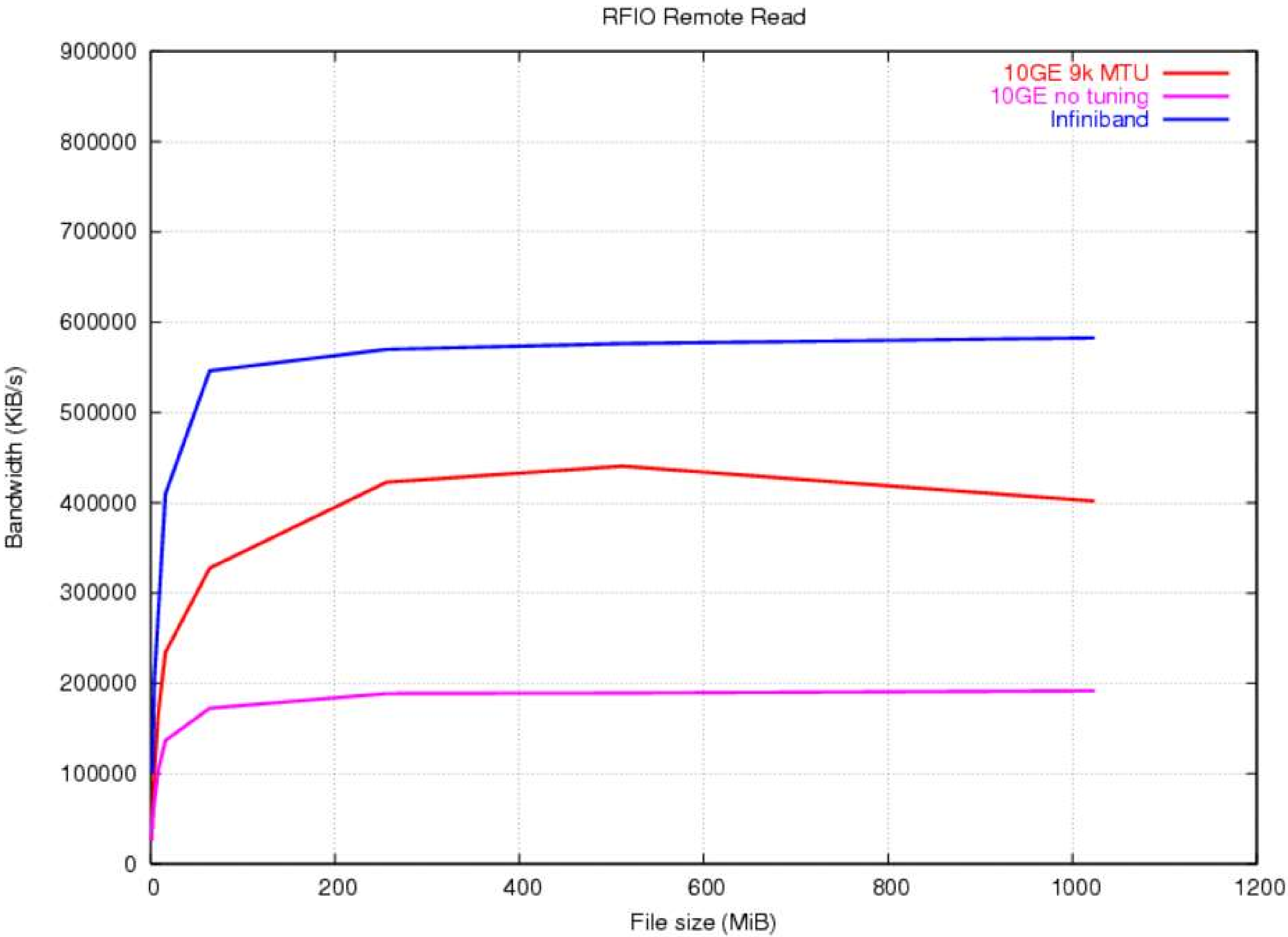



Iperf/SDP between two Xeons
Aggregate Bandwidth (KiB/s)
CPU (%)
No. of threads
IB Bandwidth
GE Bandwidth
IB Sender CPU
IB Receiver CPU
GE Sender CPU
GE Receiver CPU
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
0
20
40
60
80
100
0
5
10
15
20
25
30
35
40

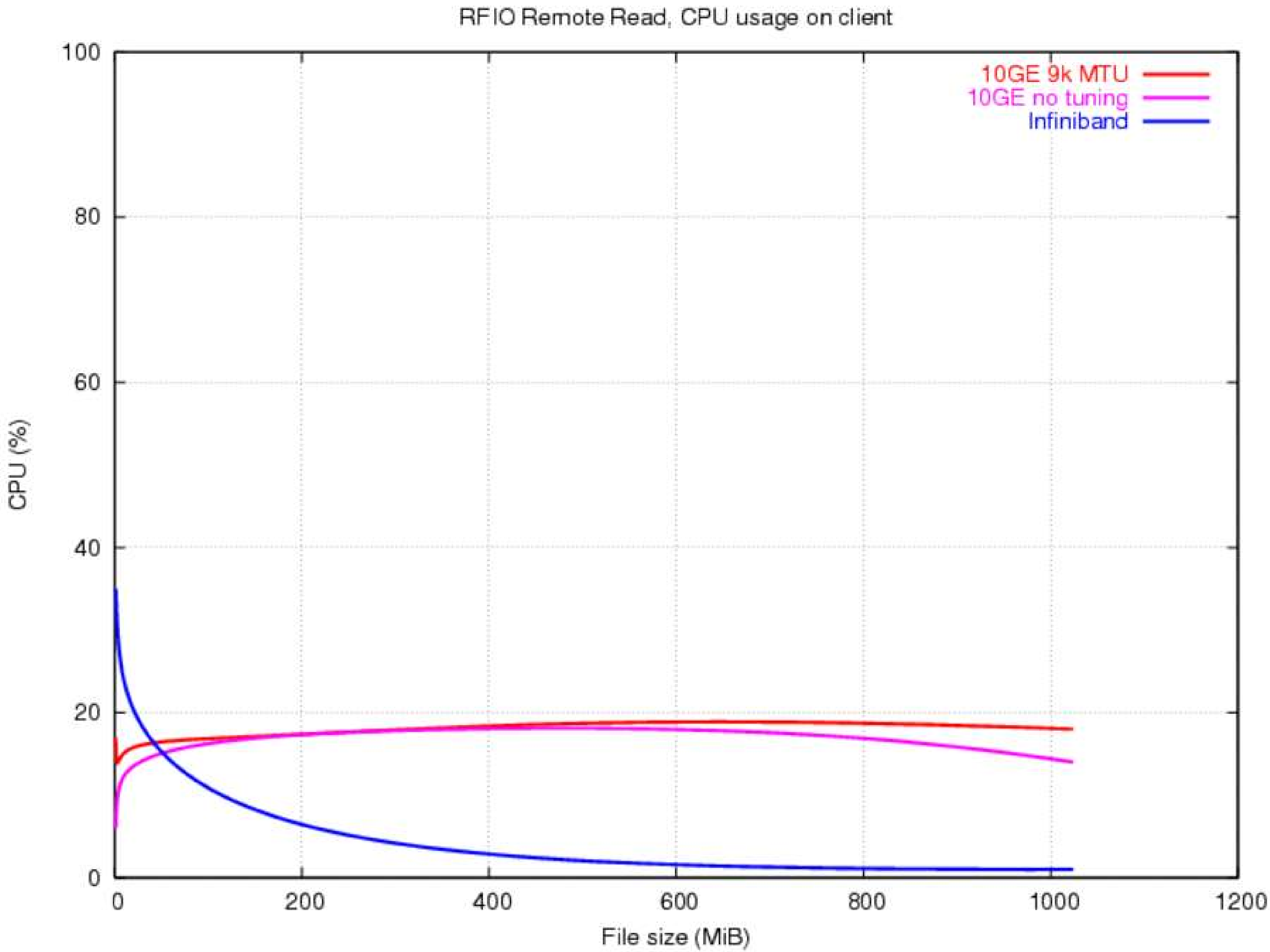
RFIO, renewed

- Code by Dr Ulrich Schwickerath, Forschungszentrum Karlsruhe
- Debug, benchmarking, functional tests at CERN (A. Horvath)
- RDMA with new streaming protocol
- Can fall back to TCP/IP
- On its way to standard Castor
- Tested on three platforms (IA32, IA64, x86_64)
- Performance tests based on IA64 tests
- dual Itanium2 1.5Ghz, 2GB RAM

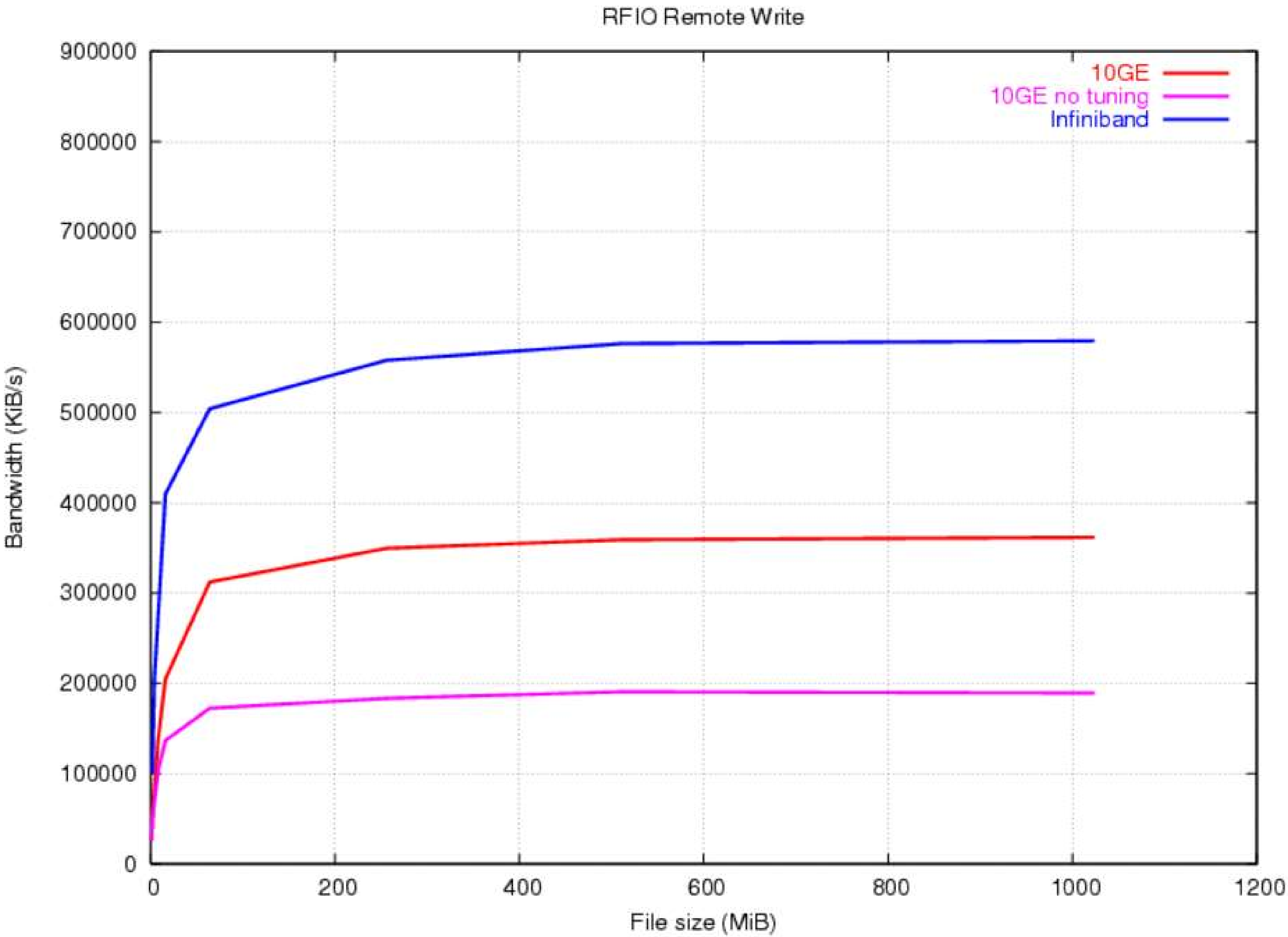


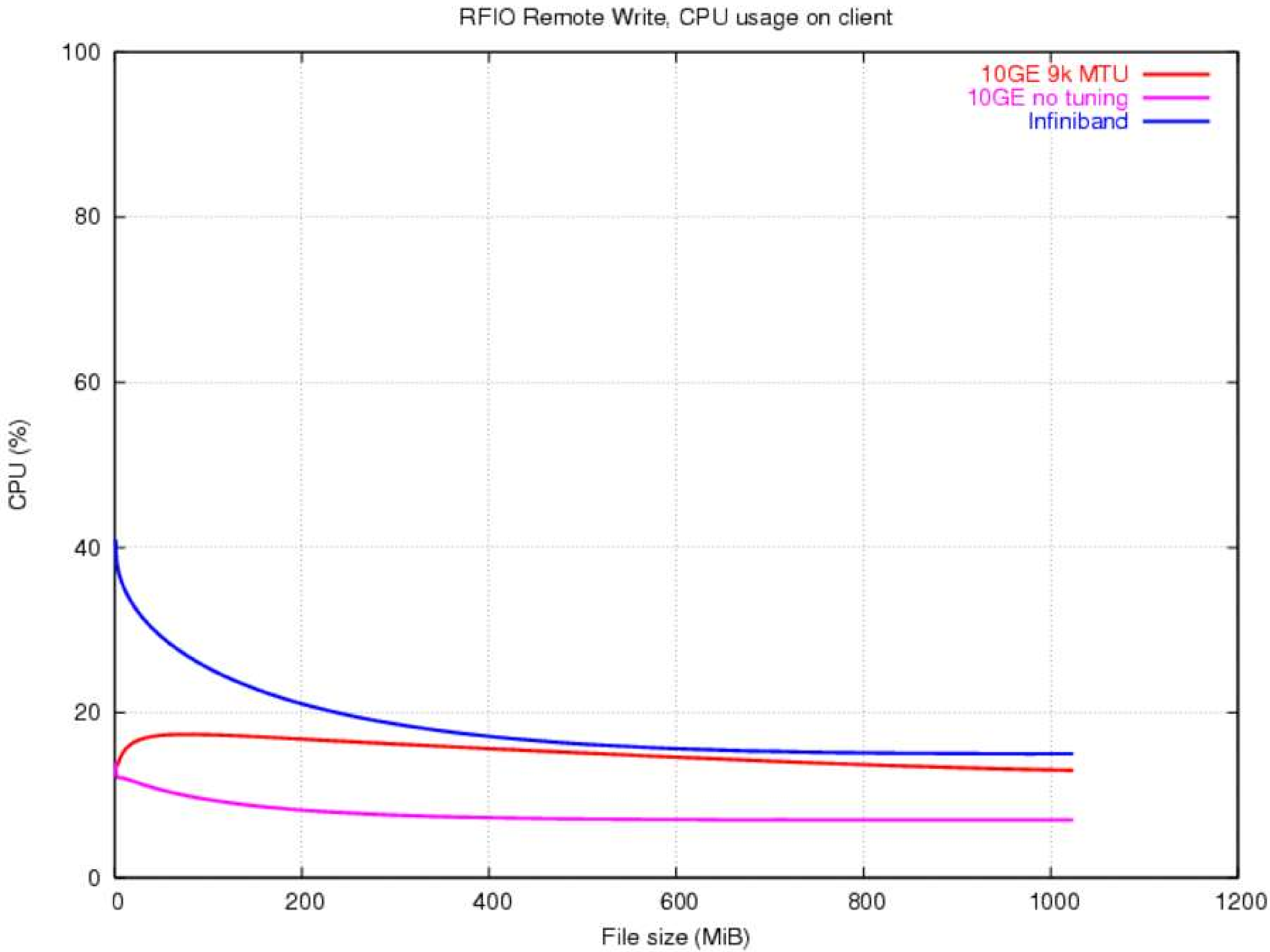


RFIO Remote Read
10GE 9k MTU
10GE no tuning
Infiniband
Bandwidth (KiB/s)
File size (MiB)
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
0
200
400
600
800
1000
1200



RFIO Remote Read, CPU usage on client
10GE 9k MTU
10GE no tuning
Infiniband
CPU (%)
File size (MiB)
0
20
40
60
80
100
0
200
400
600
800
1000
1200


RFIO Remote Write
10GE
10GE no tuning
Infiniband
Bandwidth (KiB/s)
File size (MiB)
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
0
200
400
600
800
1000
1200



RFIO Remote Write, CPU usage on client
10GE 9k MTU
10GE no tuning
Infiniband
CPU (%)
File size (MiB)
0
20
40
60
80
100
0
200
400
600
800
1000
1200

Conclusion so far

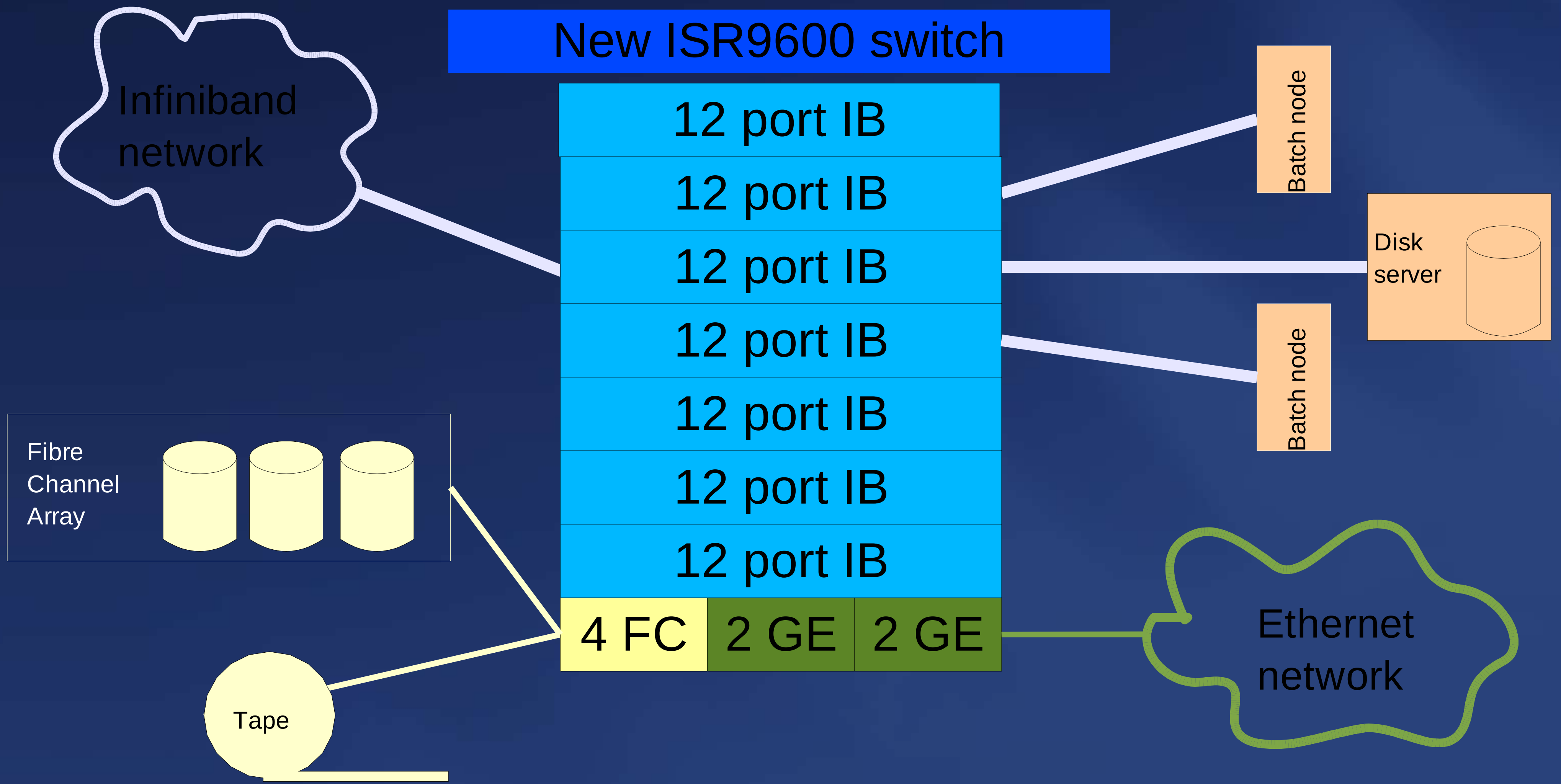
- Infiniband: good in CPU+I/O intensive environments
- market picking up (success stories)
- disruptive (cabling etc)
- applications may need porting
- still very new – prices, expertise

What next?

- Oracle 10G (on IA32 and IA64 platforms)
- Network resilience / failover testing
- Coupling to Ethernet and Fibre Channel
- Throughput (various protocols incl. ROOTd)
- Disk – to – disk transfers (Openlab efforts)
- Network filesystems?
- Anyone for MPI?

Andras Horvath

Connectivity: Fibre Channel, Ethernet



Thank you for your attention

- Generic Infiniband info: http://www.infinibandta.org
- RFIO over IB: http://www.fzk.de/infiniband/rfio.html
- CERN Openlab: http://cern.ch/openlab
- Voltaire home: http://www.voltaire.com
- Mellanox home: http://www.mellanox.com
- Opensource IB stack: http://www.openib.org

Diskserver Performance Evolution

KELEMEN PÉTER
CERN IT-ADC-LE

Overview

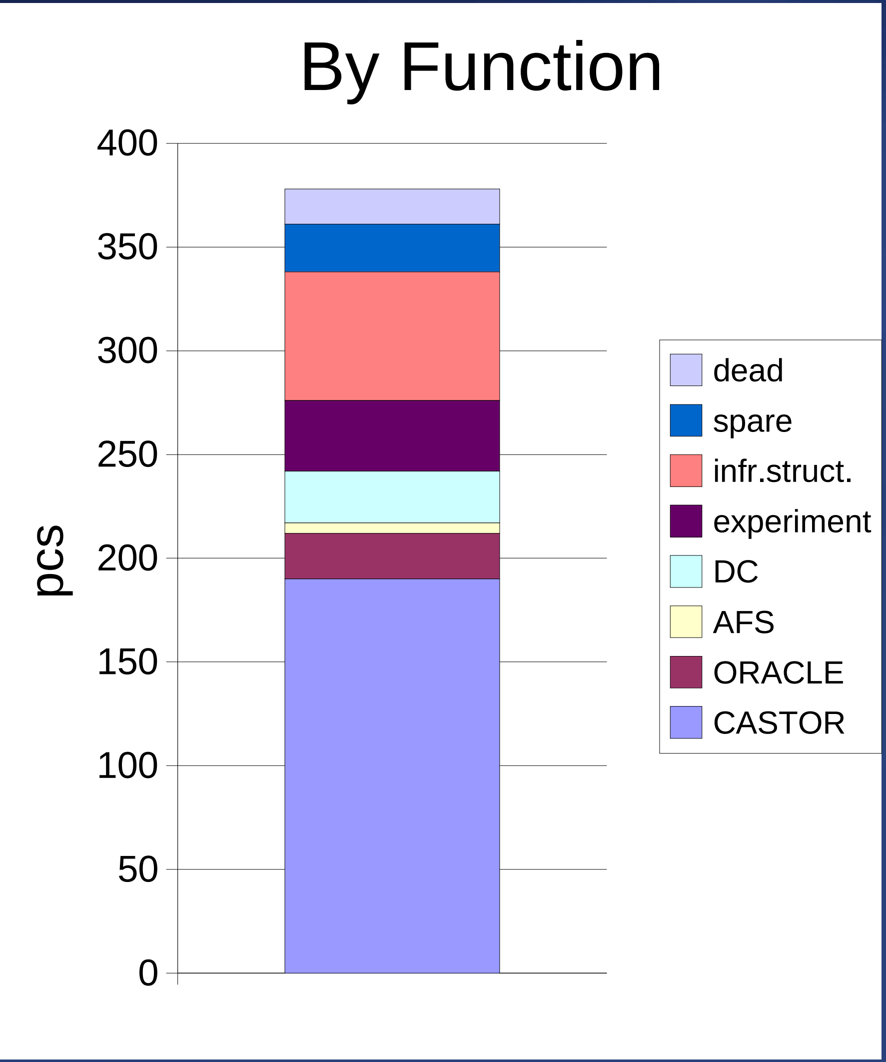
- What are diskservers?
- Generations, feature comparisons
- Possible improvements
- Performance comparisons (WRITE/READ)
- Recommendations, conclusion

Overview

- **What are diskservers?**
- Generations, feature comparisons
- Possible improvements
- Performance comparisons (WRITE/READ)
- Recommendations, conclusion

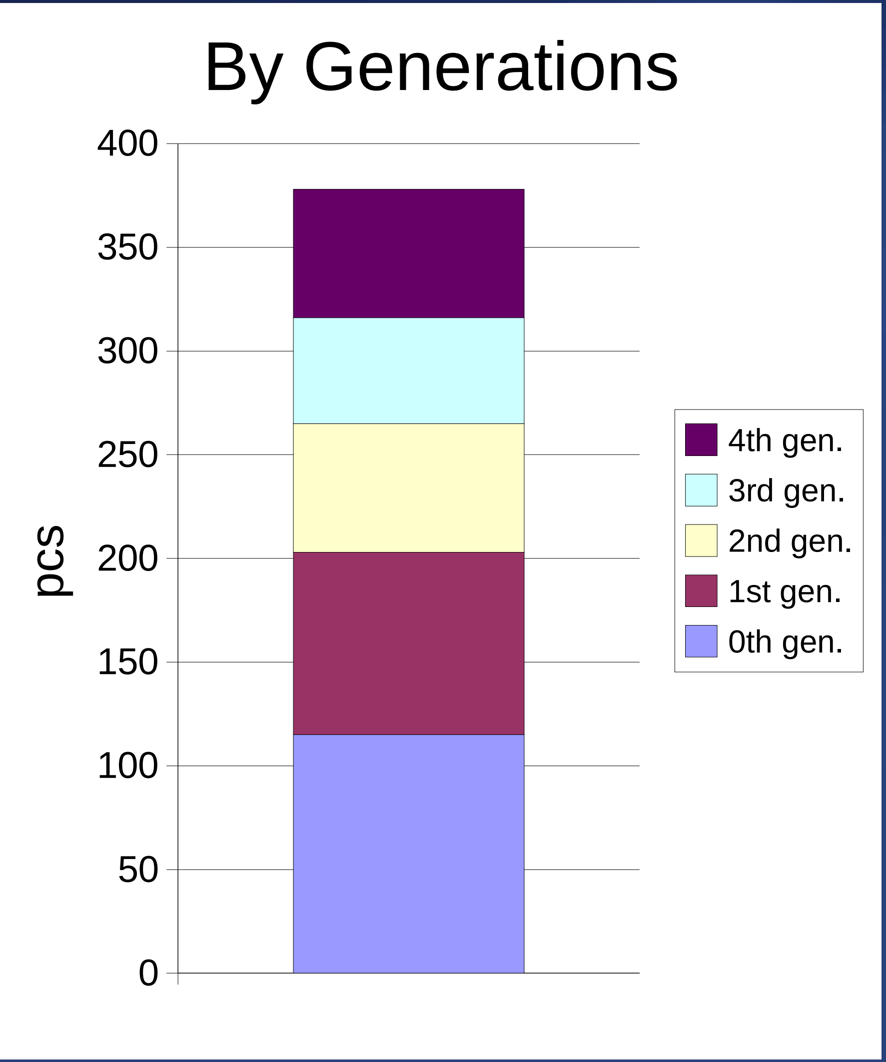
Diskservers by Function

- CASTOR stagers
- ORACLE databases
- AFS servers
- Data Challenges
- experiment-specific



Diskservers by Generation

- ~370 diskservers
- ~1TB average capacity
- Intel-based PCs
- concept from 1999
- 4 generations so far



Workloads, Redundancy

- streaming I/O vs. random I/O
- write-intensive vs. read-intensive vs. mixed
- large files vs. small files
- lots of files vs. few files
- precious data vs. throwaway (reproducible) data

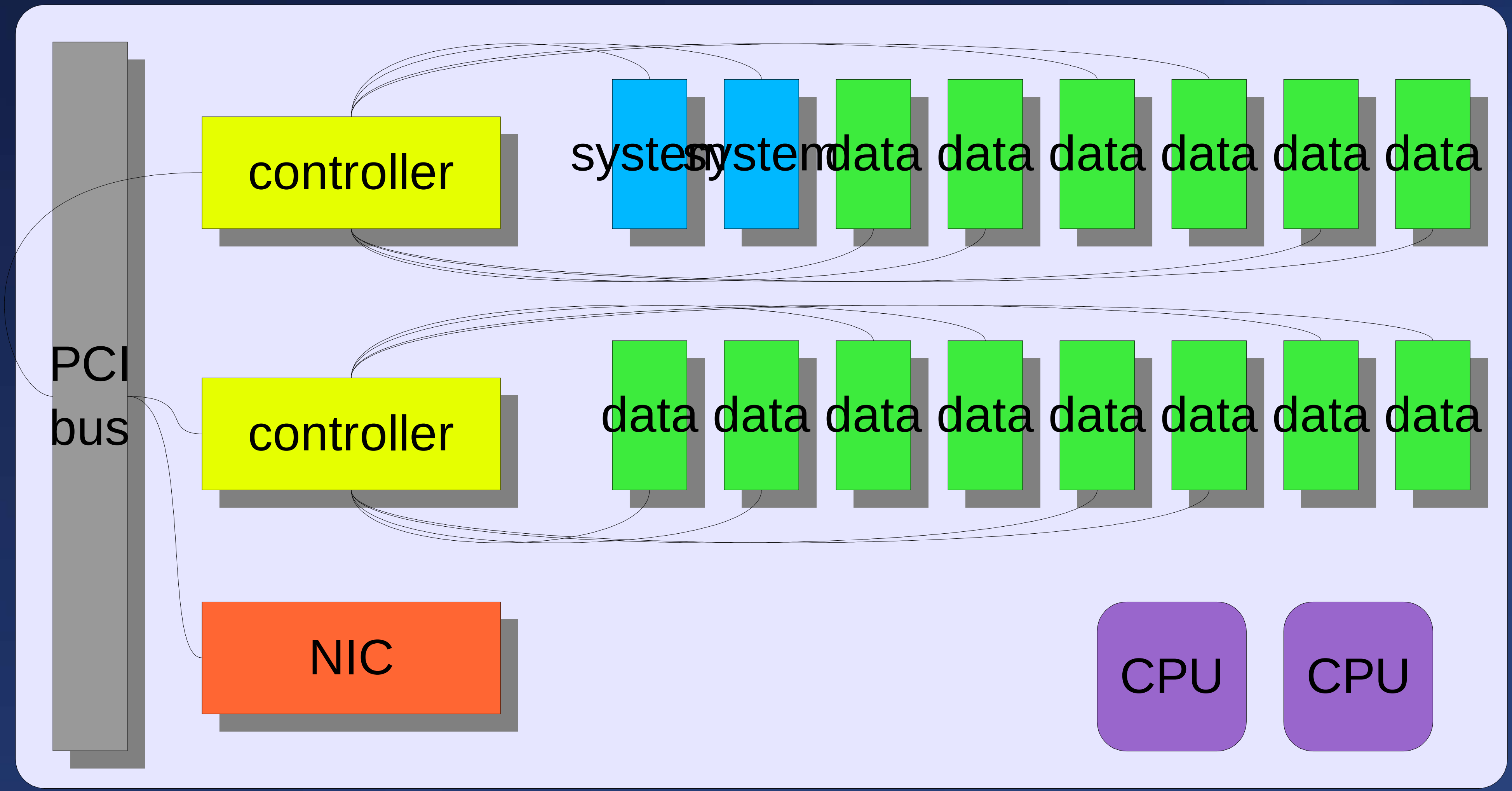
High I/O Demand

- CASTOR stagers are the major application
  - moderate number of large files
  - full sequential READ of files
  - full sequential WRITE of files
- tuning for this kind of workload makes sense
  - current filesystem: ext3 exclusively
  - current redundancy: mirrored disks exclusively

Overview

- What are diskservers?
- **Generations, feature comparisons**
- Possible improvements
- Performance comparisons (WRITE/READ)
- Recommendations, conclusion

General Diskserver Architecture



0th Generation (0G)

- 1999
- raw capacity 800GB
- usable capacity 400GB
- 20x 40GB disks
- 3x 3ware 5800-8
- dual Pentium III 650MHz
- 512M RAM
- Chieftec jumbo box
- 1.6TB/shelf

1st Generation (1G)

- April 2002
- raw capacity 1200GB
- usable capacity 600GB
- 10x 120GB WD disks
- 1x 3ware 7850-8
- 1x 3ware 7450-4
- dual Pentium III 1.13 Ghz
- 1 GB RAM
- 7.2 TB/rack (5U), 9.6 TB/rack (4U)

2^{nd} Generation (2G)

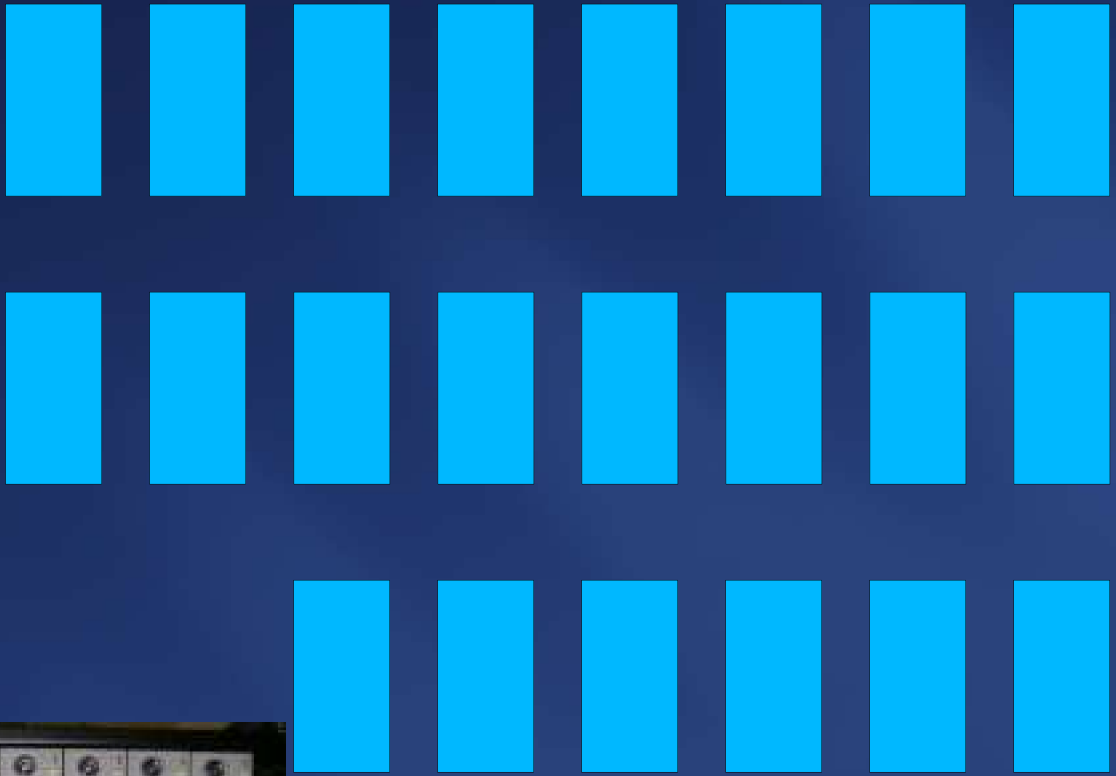
- October 2002
- raw capacity 1440GB
- usable capacity 720GB
- 12x 120GB WD disks
- 2x 3ware 7500-8
- dual Xeon 2.0GHz
- 1 GB RAM
- 11.52 TB/rack

3^{rd} Generation (3G)

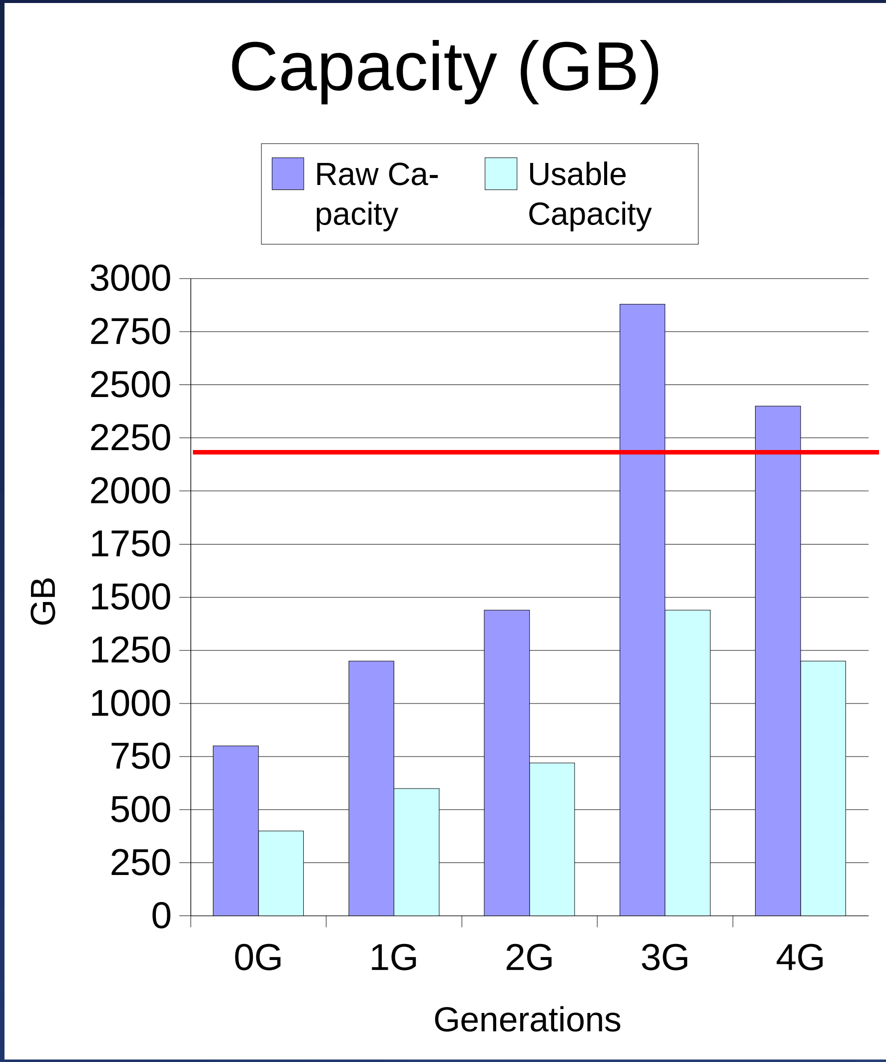
- March 2003
- raw capacity 2880GB
- usable capacity 1440GB
- 24x 120GB WD disks
- 4x 3ware 7500-8
- dual Xeon 2.4GHz
- 1 GB RAM
- 11.52 TB/rack

4th Generation (4G)

- October 2003
- raw capacity 2400GB
- usable capacity 1200GB
- 20x 120GB WD disks
- 3x 3ware 7506-8
- dual Xeon 2.4GHz
- 2 GB RAM
- 9.6 TB/rack

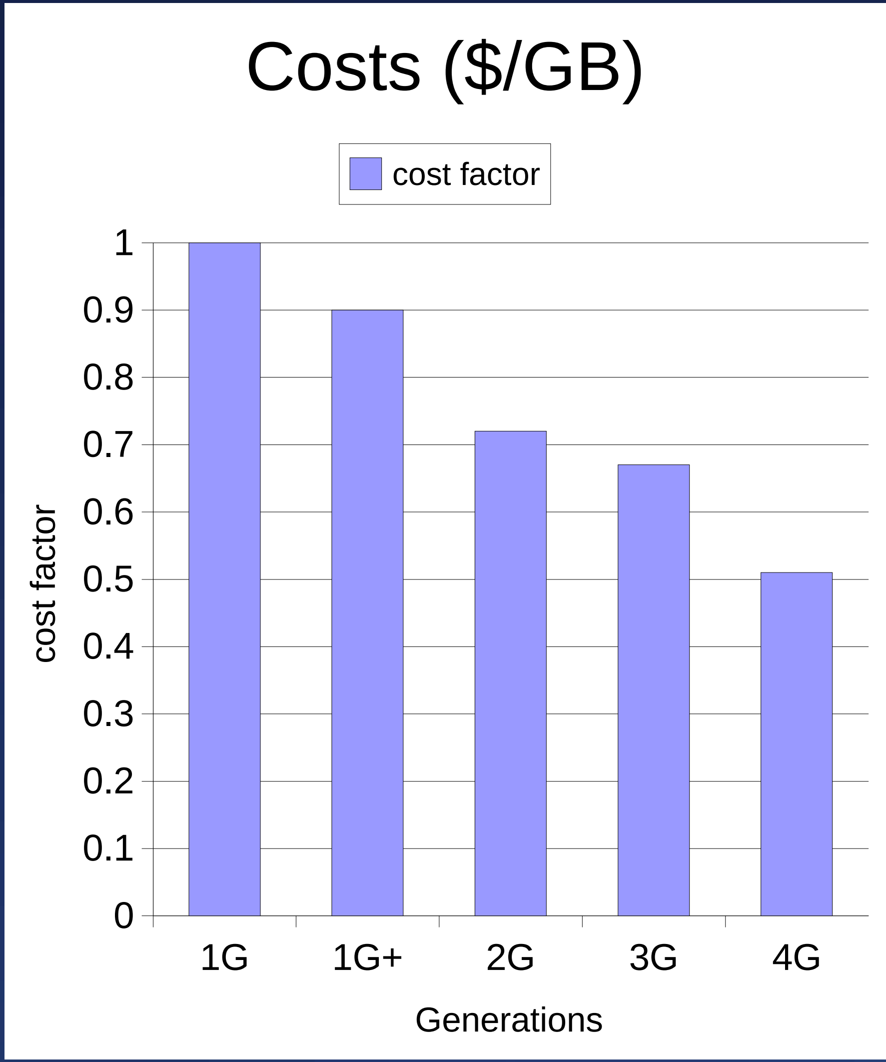


Capacity Comparison



- |3G| / |0G| = 3.6
- Linux 2.4.x max. block device size 2 TiB
- 1 GB ~ 0.9313 GiB
- 2199 GB ~ 2 TiB

Costs Comparison

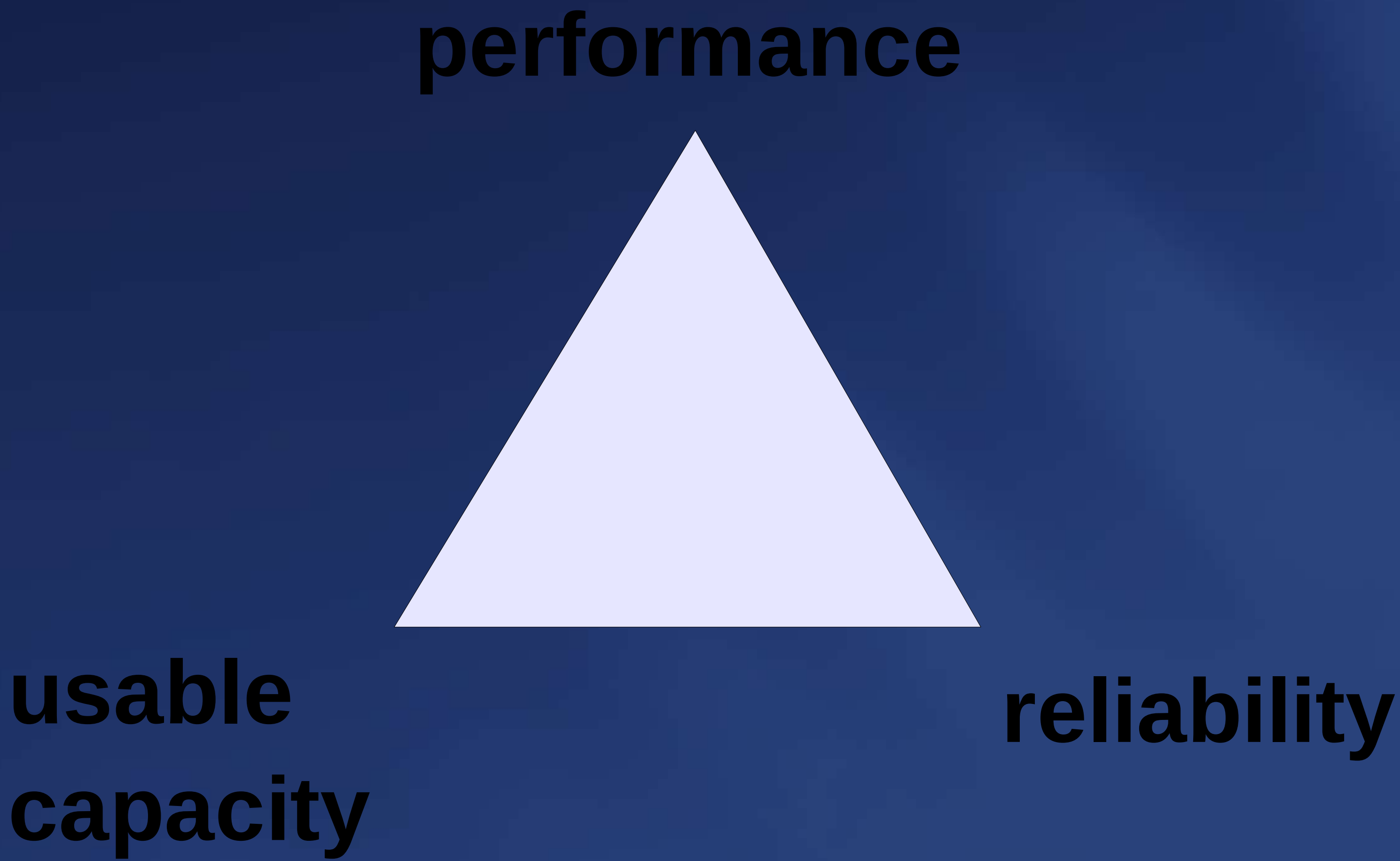


- exact numbers are not available due to confidentiality reasons
- numbers are provided by FIO

Overview

- What are diskservers?
- Generations, feature comparisons
- **Possible improvements**
- Performance comparisons (WRITE/READ)
- Recommendations, conclusion

Trade-Offs

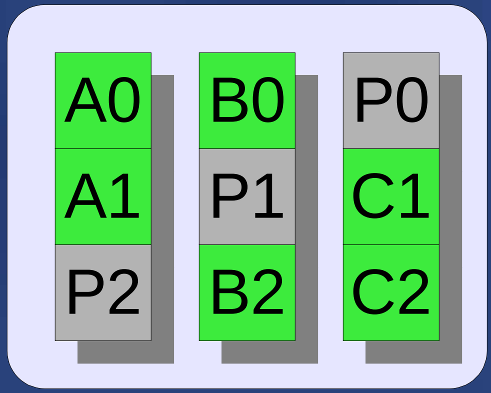


Degrees of Freedom

- RAID configurations
  - performance
  - usable capacity
  - reliability
- filesystems
  - performance
- kernel tuning
  - performance

RAID Primer

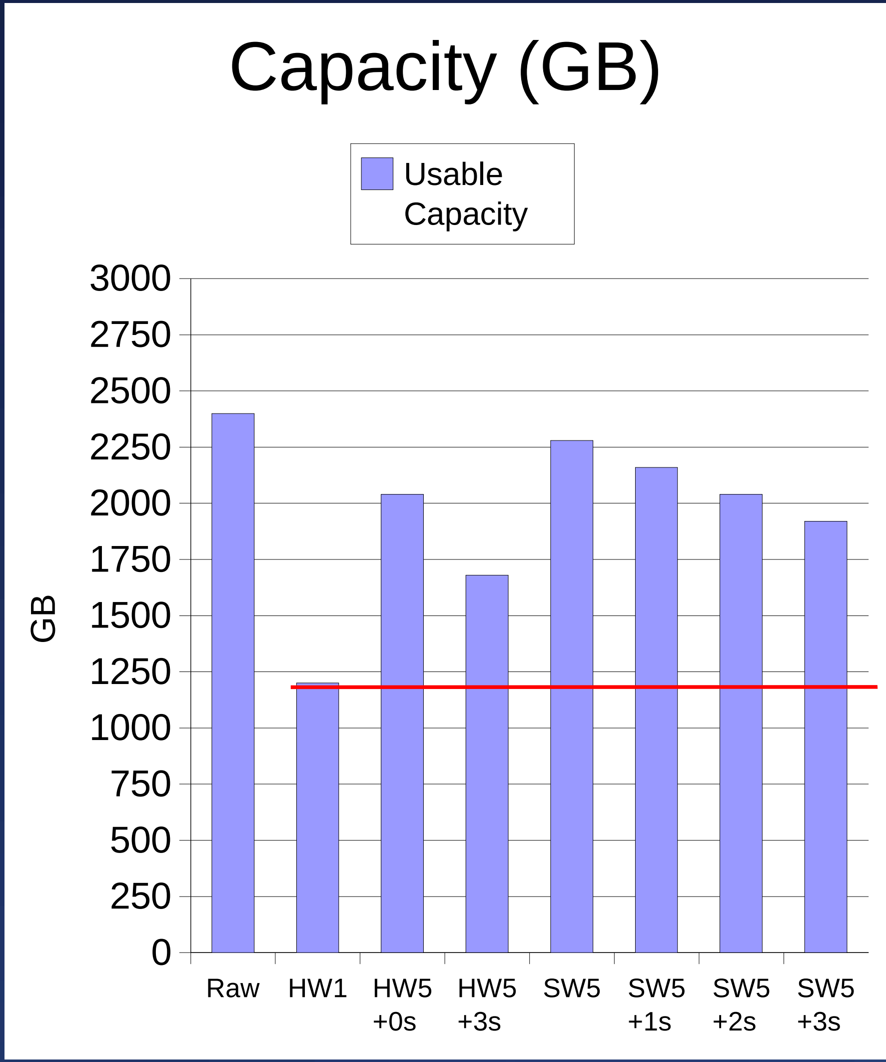
- Redundant Array of Independent Disks
  - data stored in multiple places to achieve increased fault-tolerance (MTBF) and/or load-balancing
- several levels defined in the standards
  - RAID-0 (striping)
  - RAID-1 (mirroring)
  - RAID-5 (rotating parity)
  - ...combinations
- implementation
  - hardware (RAID controllers)
  - software (kernel modules)



RAID Setups

- “current”: disks mirrored by hardware
- “HW1”: hardware mirrors
- “HW10”: hardware mirrors, hardware striping
- “HW1-SW0”: hardware mirrors, software striping
- “HW5”: hardware RAID-5
- “HW5-SW0”: hardware RAID-5, software striping
- “SW10”: software mirrors, software striping
- “SW5”: software RAID-5
- “SW50”: software RAID-5, software striping

RAID Capacity Impact (4G)



- space more than HW1:
  - HW5: 70% more
  - HW5+3s: 40% more
  - SW5: 90% more
  - SW5+1s: 80% more
  - SW5+2s: 70% more
  - SW5+3s: 60% more

Filesystems

- ext3 (RedHat default)
  - excellent track record
  - scalability problems
- XFS (SGI)
  - mature and full-featured
- JFS (IBM)
  - very promising but yet incomplete implementation
- ReiserFS (Namesys)
  - v3.6: fragile, instable fsck(8) utility
  - v4: alpha quality

Kernel Internals

- filesystem mount options
  - larger in-core journal
- VM subsystem
  - I/O balancing
  - read-ahead
- I/O scheduler (elevator)
  - I/O merging
  - latency vs. throughput
- scheduler
  - IRQ balancing

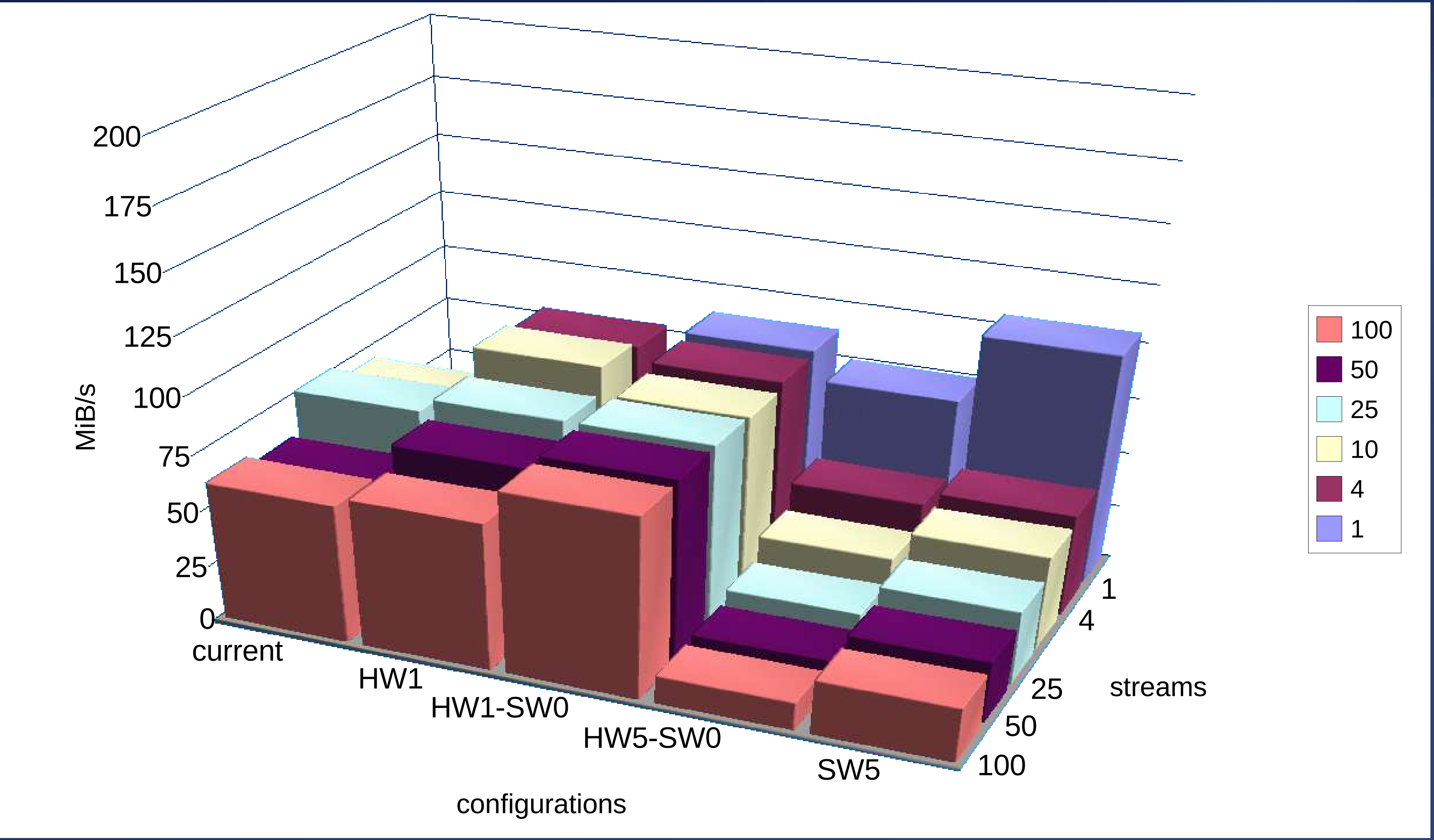
Tuning

- filesystem: XFS
  - appropriate parameters for various RAID arrays...
- kernel: elevator tuning
  - READ: 512 instead of 64
  - WRITE: 1024 instead of 8192
- kernel: VM-tuning
  - `vm.bdflush = 2 500 0 0 500 1000 60 20 0`
  - `vm.max-readahead = 256`
  - `vm.min-readahead = 127`
- all parameters carefully derived from experience, source code analysis and experimentation

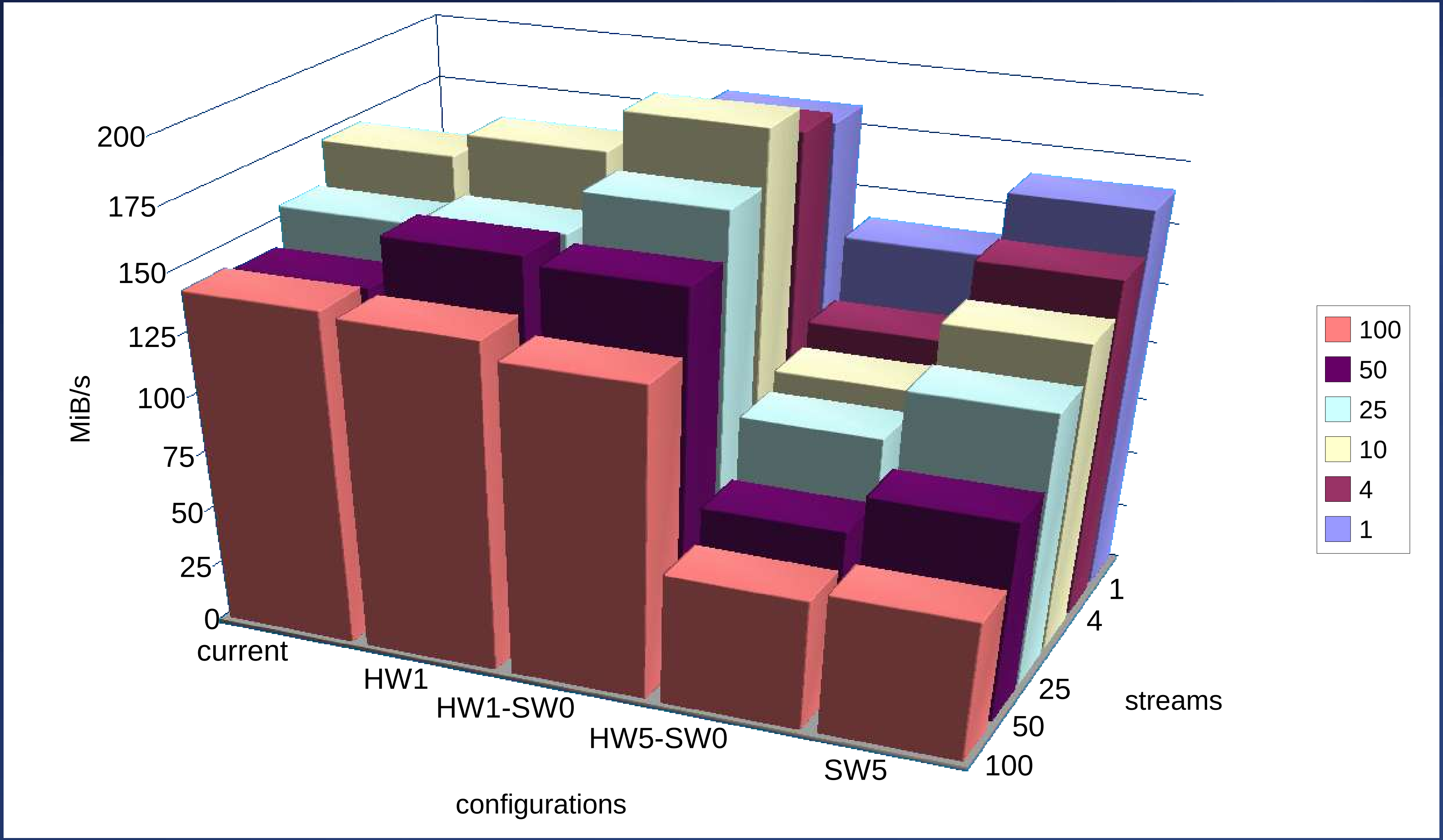
Overview

- What are diskservers?
- Generations, feature comparisons
- Possible improvements
- **Performance comparisons (WRITE/READ)**
- Recommendations, conclusion

2G WRITE Comparison

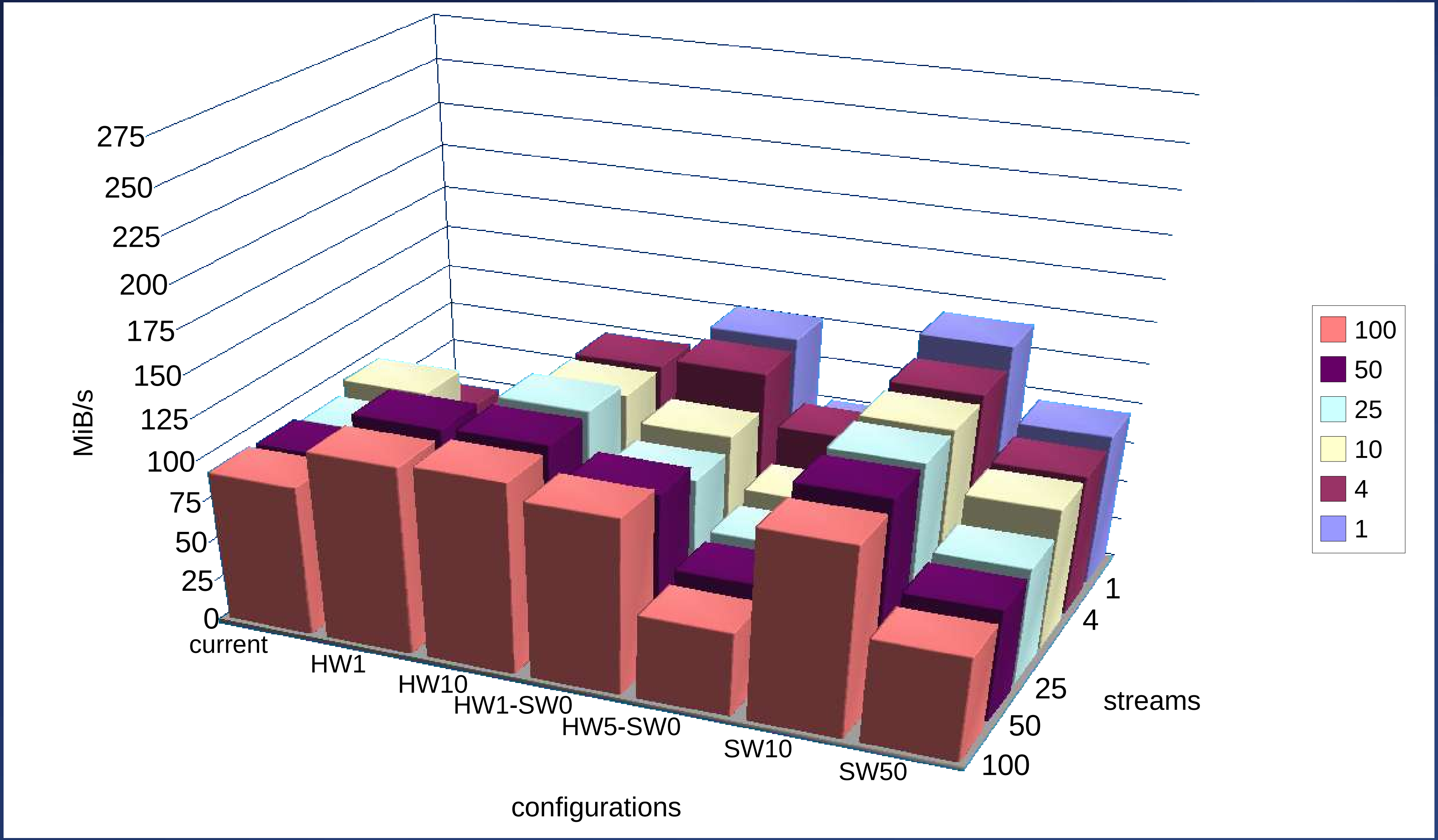


2G READ Comparison



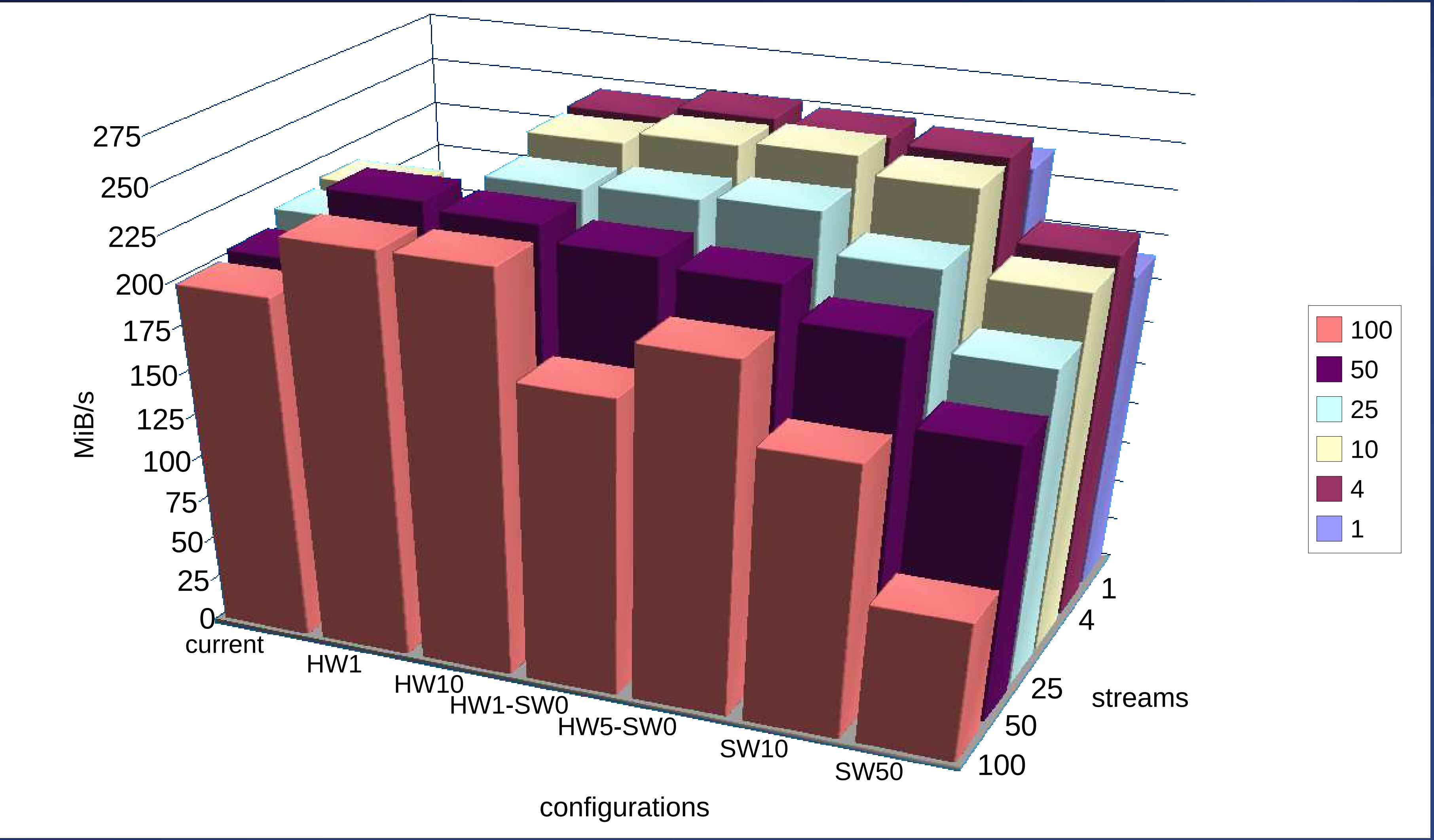
Péter Kelemen

3G WRITE Comparison

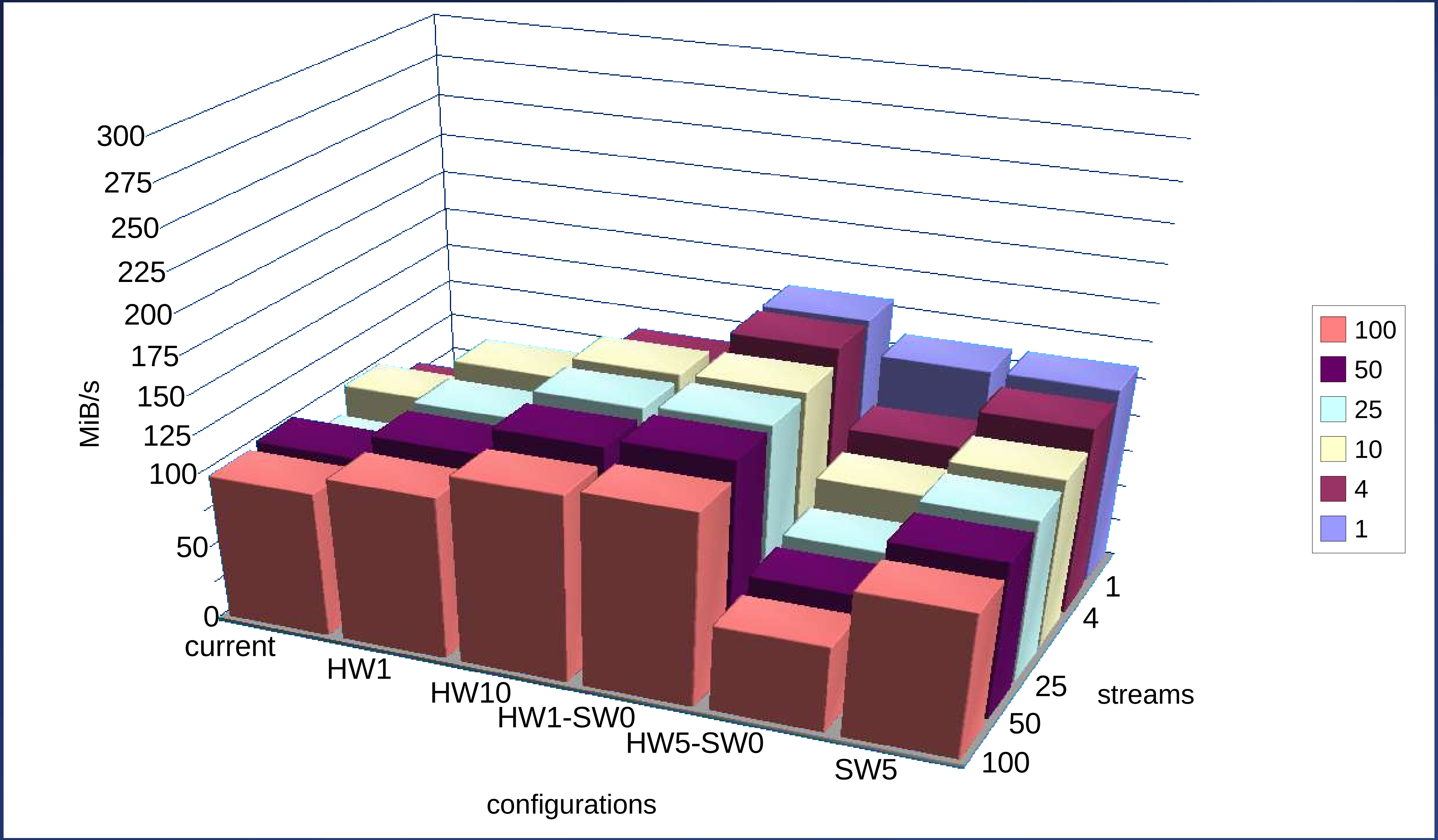


Péter Kelemen

3G READ Comparison

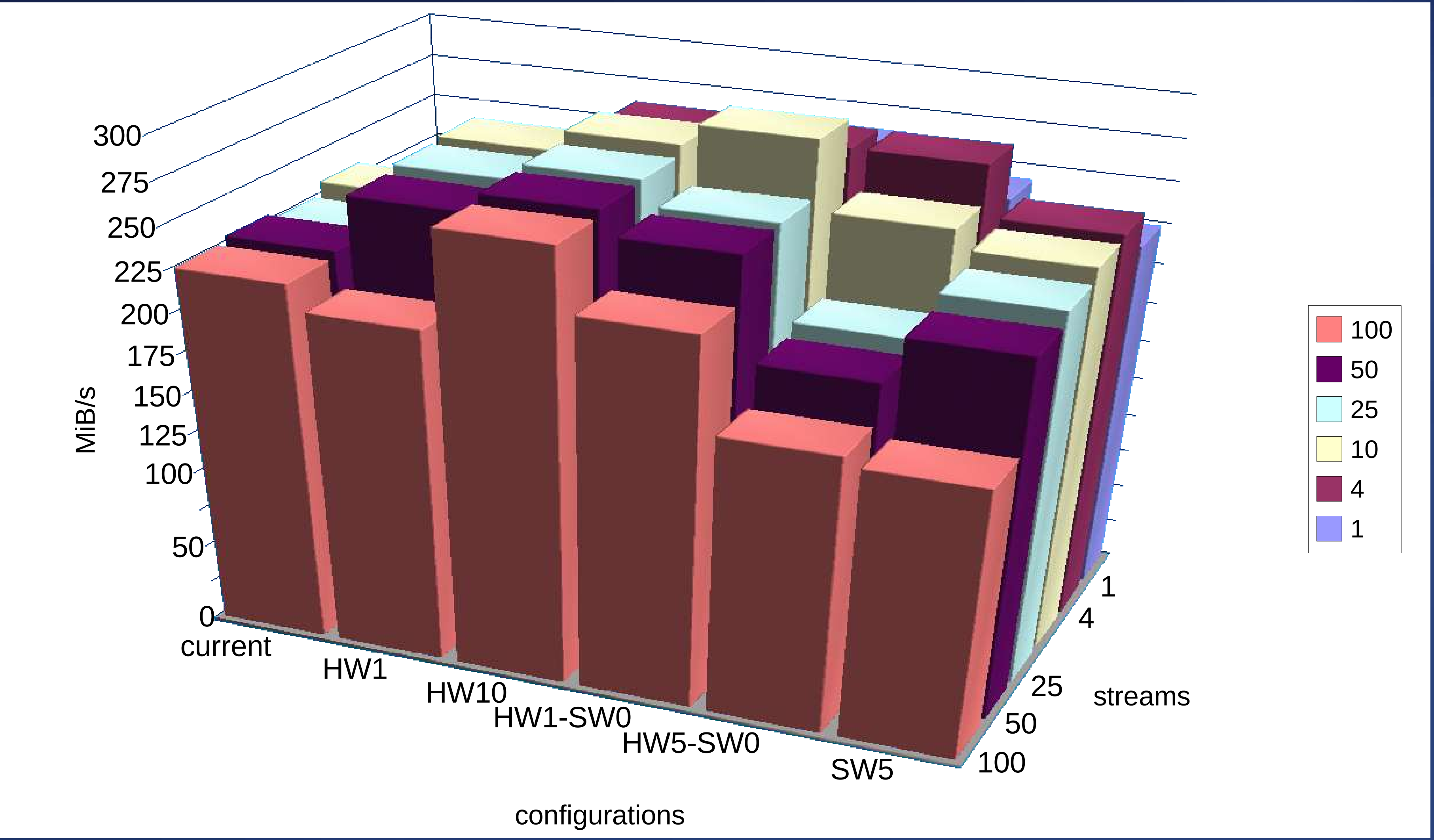


4G WRITE Comparison



Péter Kelemen

4G READ Comparison



Overview

- What are diskservers?
- Generations, feature comparisons
- Possible improvements
- Performance comparisons (WRITE/READ)
- **Recommendations, conclusion**

Recommendations

- WRITE performance since READ is good enough
- Goals
    - same performance with increased capacity

      OR
    - increased performance with same capacity
- 2G: HW1-SW0 for perf, SW5 for capacity
- 3G: HW1-SW0 for perf, SW5 for capacity
- 4G: HW1-SW0 for perf, SW5 for capacity
- XFS as filesystem
- appropriate kernel tuning

Feasibility

- CERN Linux 7.3.x is OK with additional packages
- CEL3 is OK out of the box
- automated changes available
  - hardware RAID configurations
  - software RAID configurations
  - filesystem
  - kernel tuning
- hooks into FIO procedures

Conclusions

- no “silver bullet”
- it is crucial to know your workload
- significant improvements are possible in many directions (performance, capacity)
- practically zero cost
- incremental transition is relatively easy
- implementation can start tomorrow
- regular re-evaluation is highly recommended