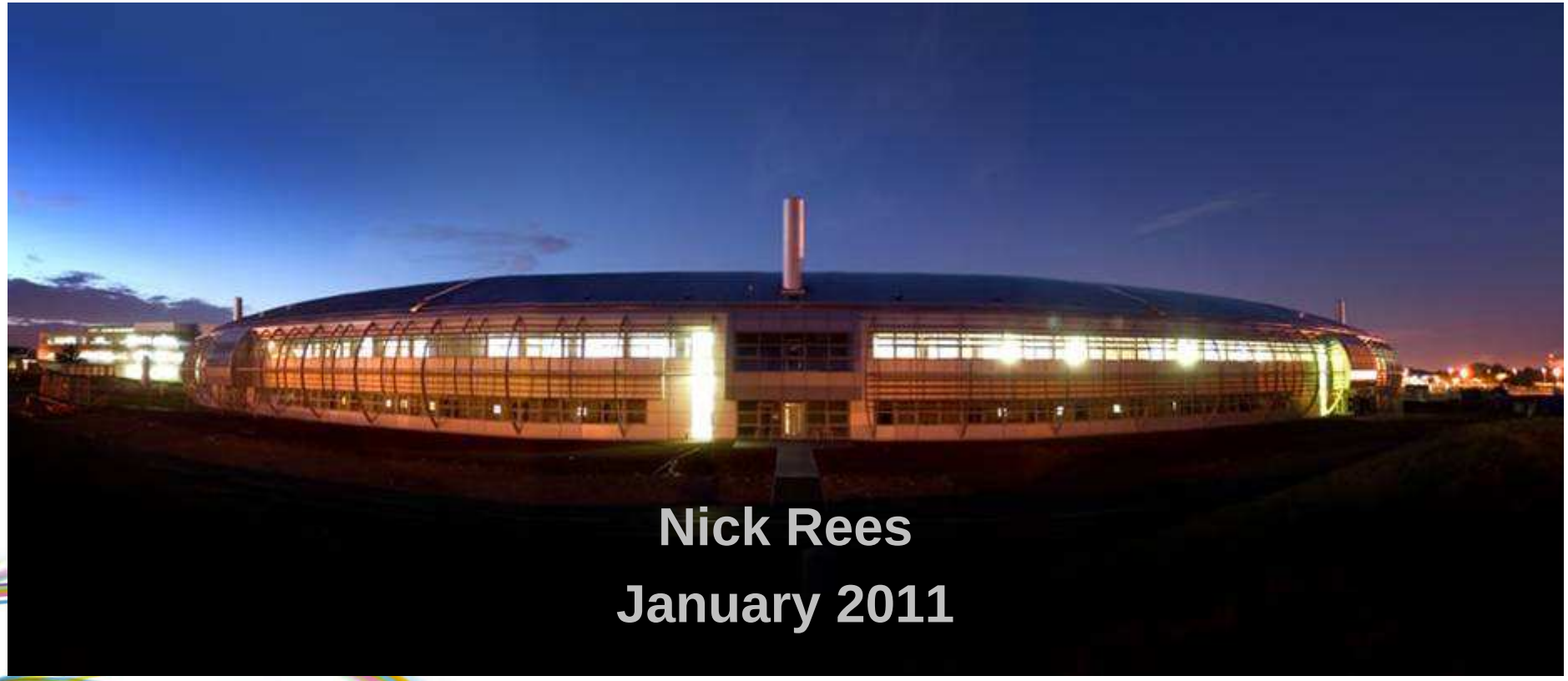
Diamond Networks/Computing



Nick Rees

January 2011

2008 computing requirements

- Diamond originally had no provision for central science computing.
- Started to develop in 2007-2008, with a major development in 2008:
  - Resilient high density computer room.
  - Compute cluster with minimum 1500 SPECfp_rate2006 total.
  - Resilient network with dual 10 Gbit core.
  - 200 Tbytes storage with 20 Mbytes/sec scalable aggregated throughput and 8000 metadata operations/sec.

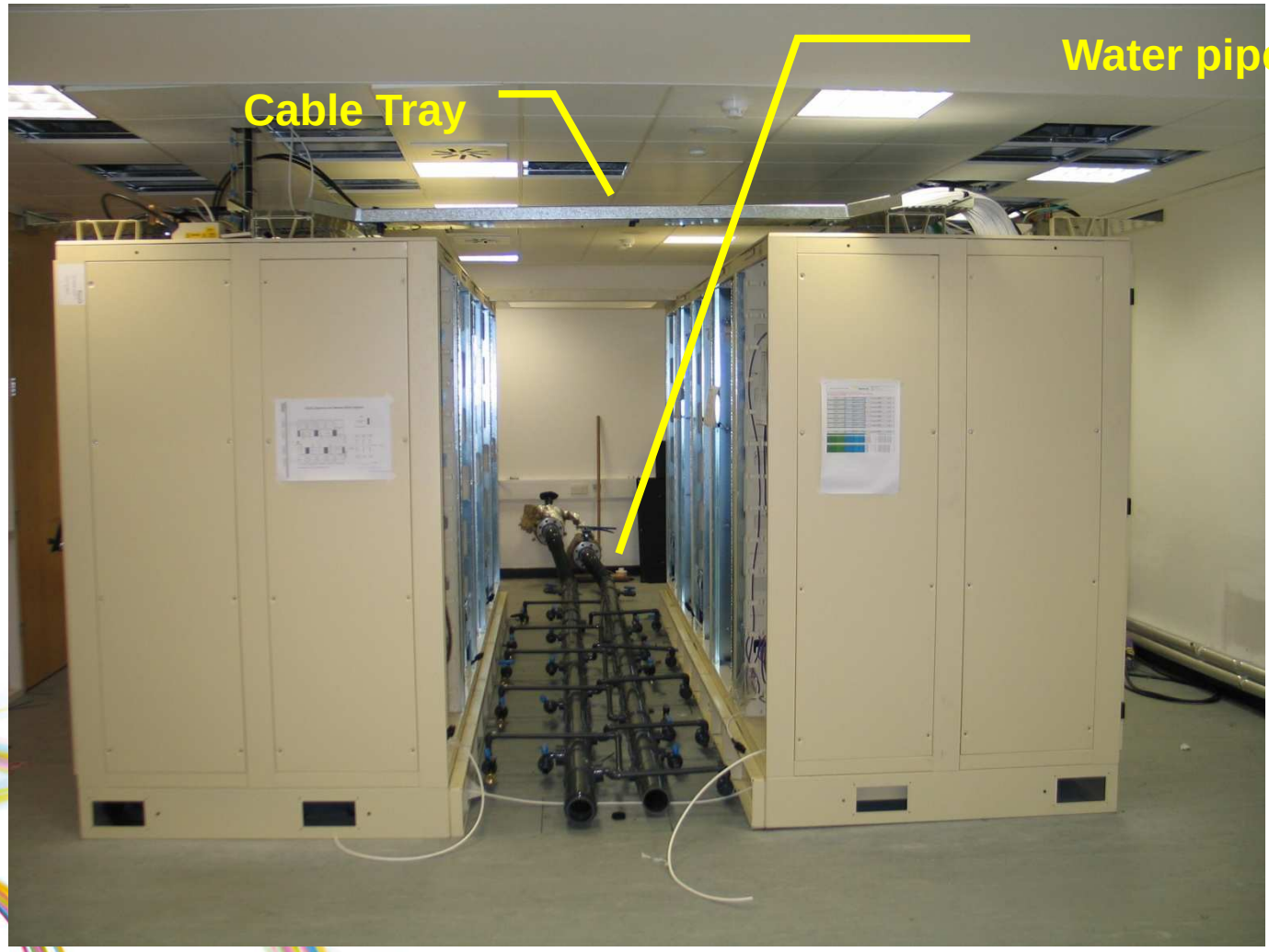
diamond

Computer room

- Overall a success
- Has up to 320 kW redundant power, (A and B) from two separate sub-stations.
- Power from A is UPS and generator backed up.
- Has up to 320 kW cooling water.
- Primary cooling is from site chilled water,
- 220 kW standby chiller in case of problems.
- Ran commissioning tests at 160 kW power load satisfactorily.
- Standby system has proved its worth a number of times.

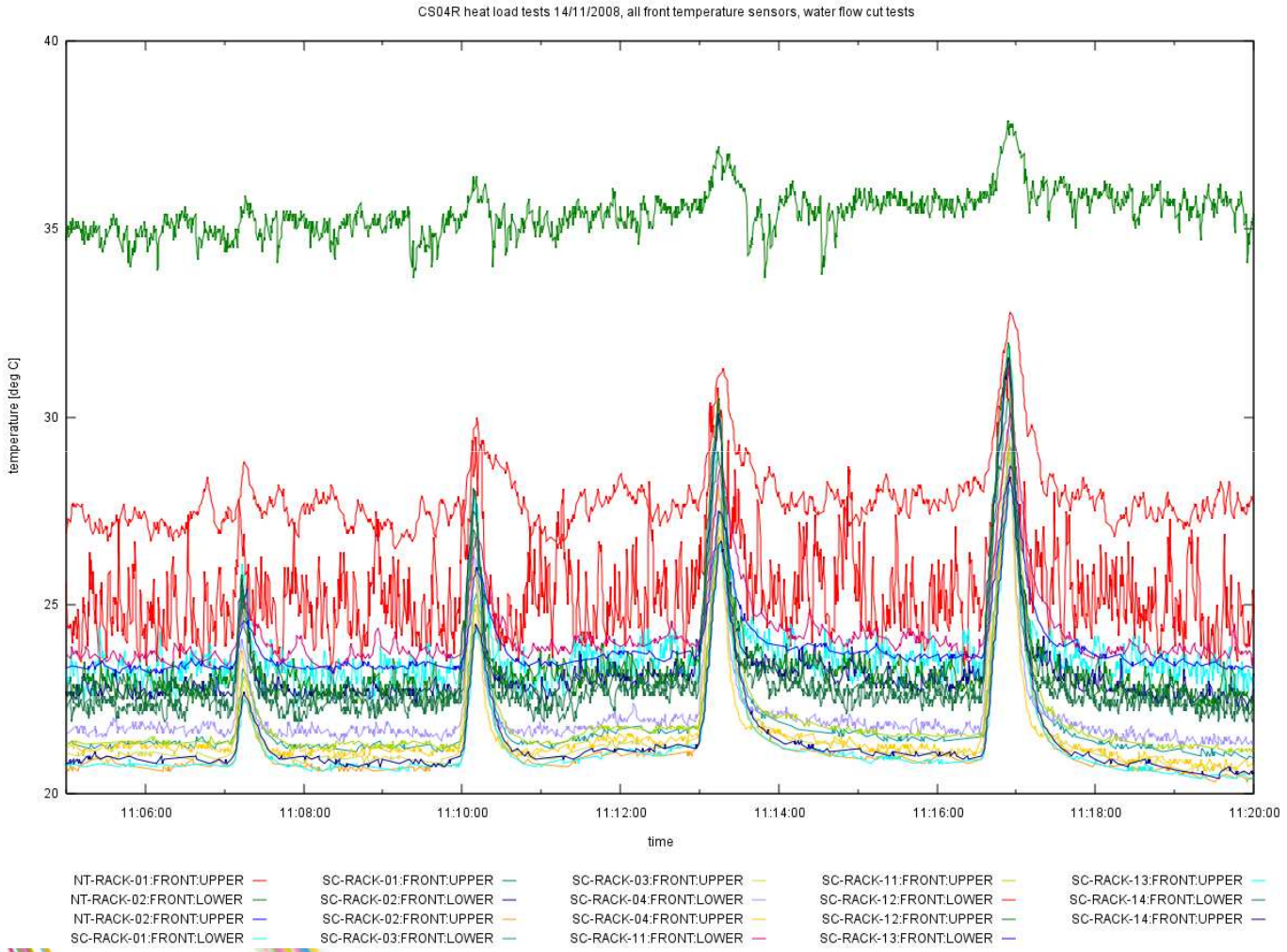
diamond

Layout

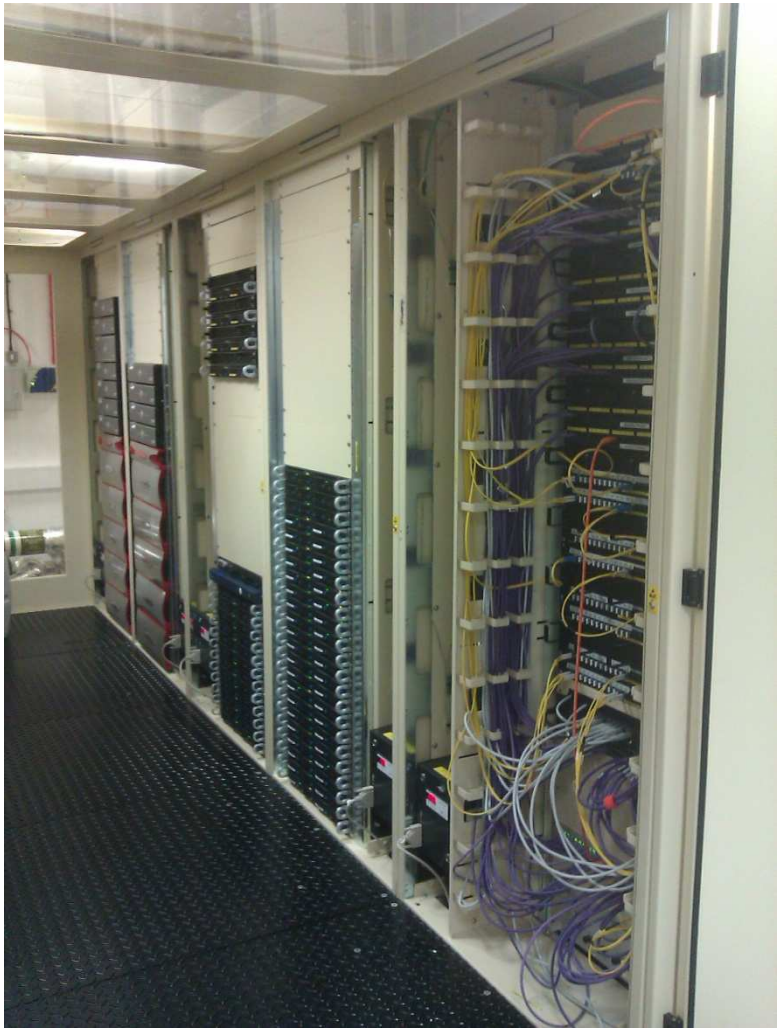


Testing with heat loads

Temperature rise on flow fail (0.5C/sec)

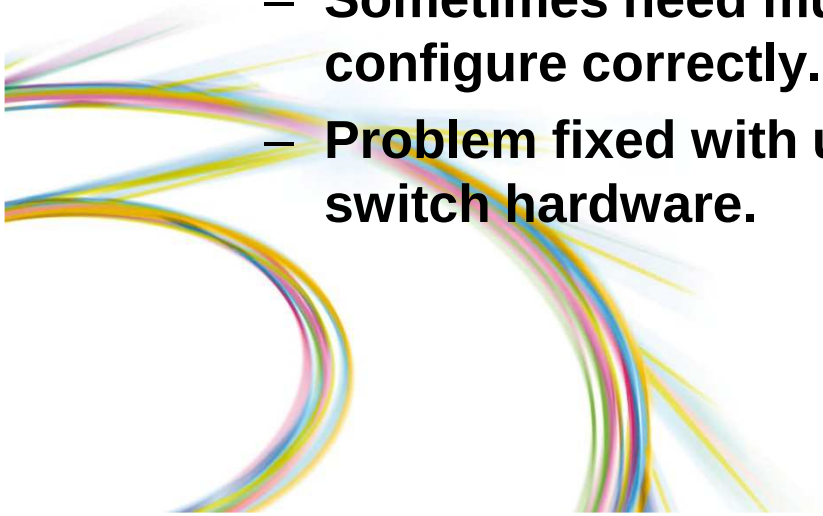


Now



Compute cluster

- This was a success.
- Easy to specify
- One supplier validated system against specification.
  - They were the successful supplier
- A few teething problems
  - Network driver had problems being both IPMI and normal interface at same time
  - Sometimes need multiple restarts to get network I/F configure correctly.
  - Problem fixed with updated Ethernet firmware and new switch hardware.



diamond

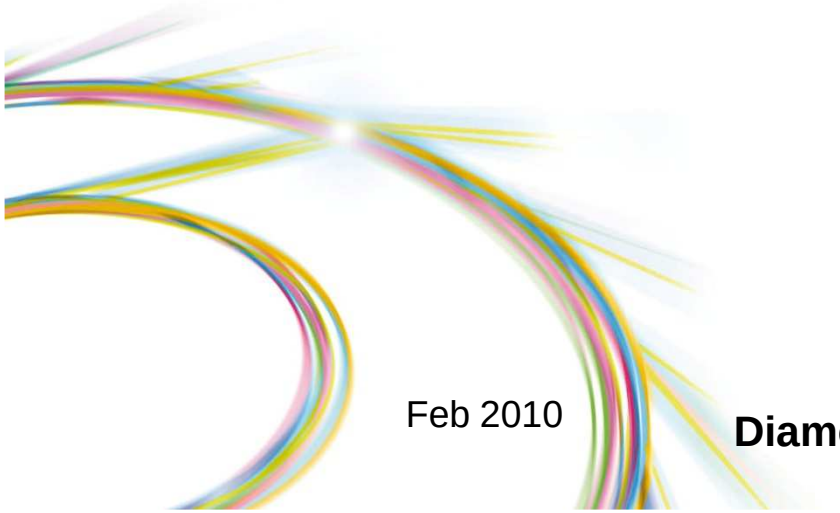
Computing – solution

- 1U twin motherboard systems
- Each 1U motherboard is:
  - Supermicro X7DWT
  - Intel 5400 (Seaburg) Chipset
  - Dual Intel Xeon E5420 (Quad Core, 2.5GHz) Processors
  - 16GB DDR2-667 ECC RAM
  - Installed as 4x4GB FBDIMMs
  - 160GB SATA Hard Disk
  - 1x16 PCI-Express slot (for GPU extension)
- Total 20 systems, 16 cores/system, 320 cores.

diamond

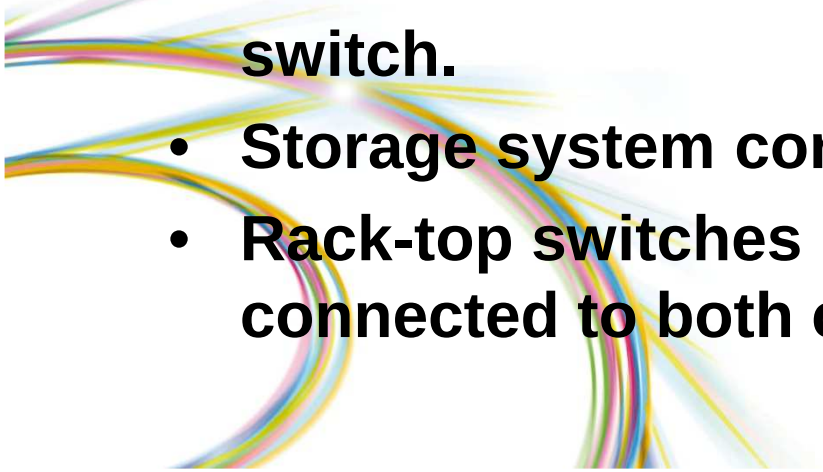
Current compute clusters

- **Now have multiple compute clusters:**
  - **Original IBM MX cluster**
  - **New Supermicro cluster, 320 cores**
  - **Accelerator Physics cluster, 240 Supermicro cores, but with InfiniBand interconnect.**
  - **Nvidia Tesla cluster with Supermicro frontends for tomography.**



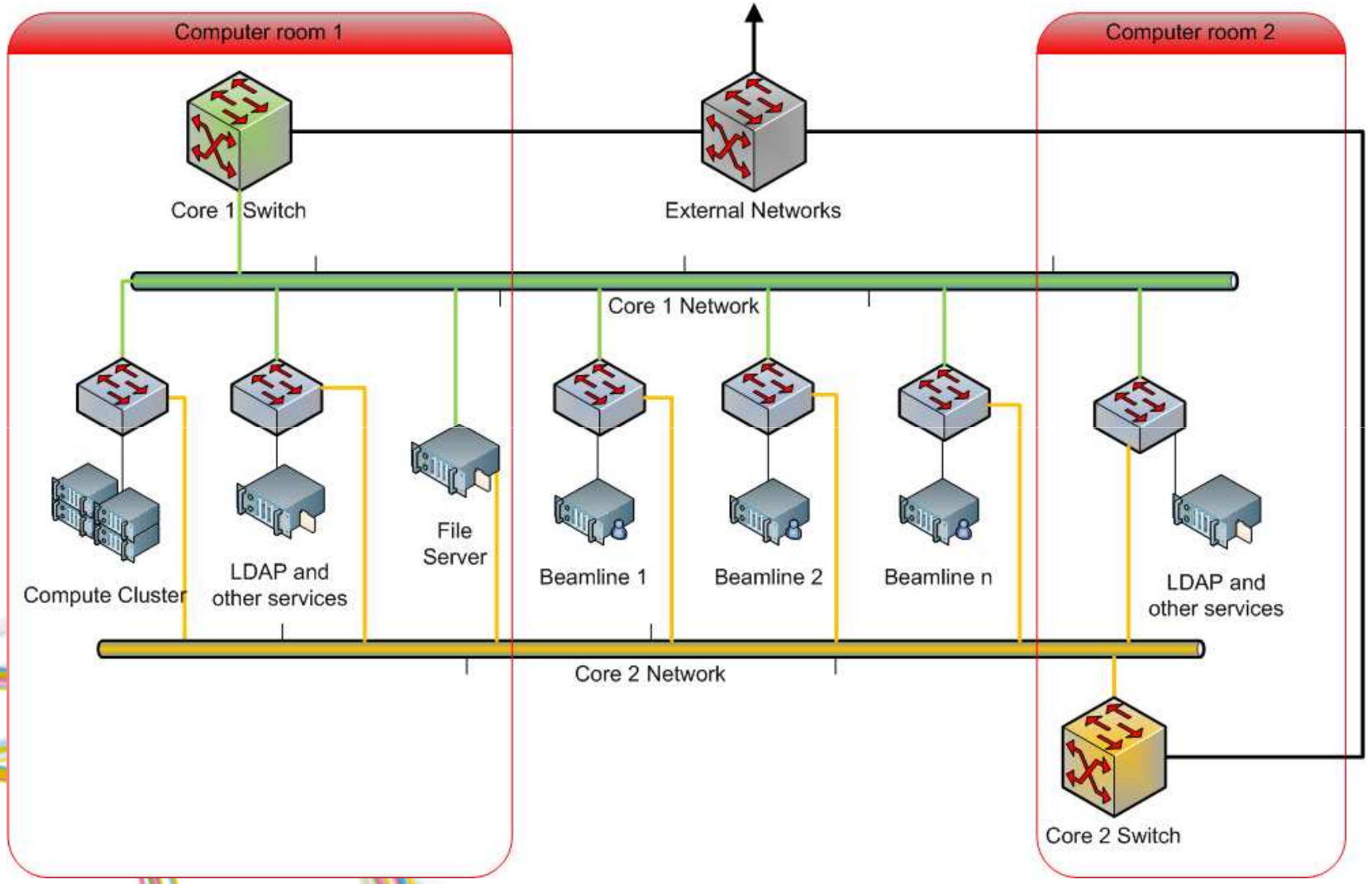
Network

- Overall, another success
- Two core switches (from different vendors), in separate computer rooms
- Each beamline connected to both switches – so either 2GBit or 20 GBit peak bandwidth available
  - halved if one core switch fails
- Traffic routed by vendor-independent protocols
  - OSPF, ECMP and LACP
- Cluster switch connected with 2 x 10 GBit to each switch.
- Storage system connected directly into core.
- Rack-top switches in both computer rooms also connected to both cores.



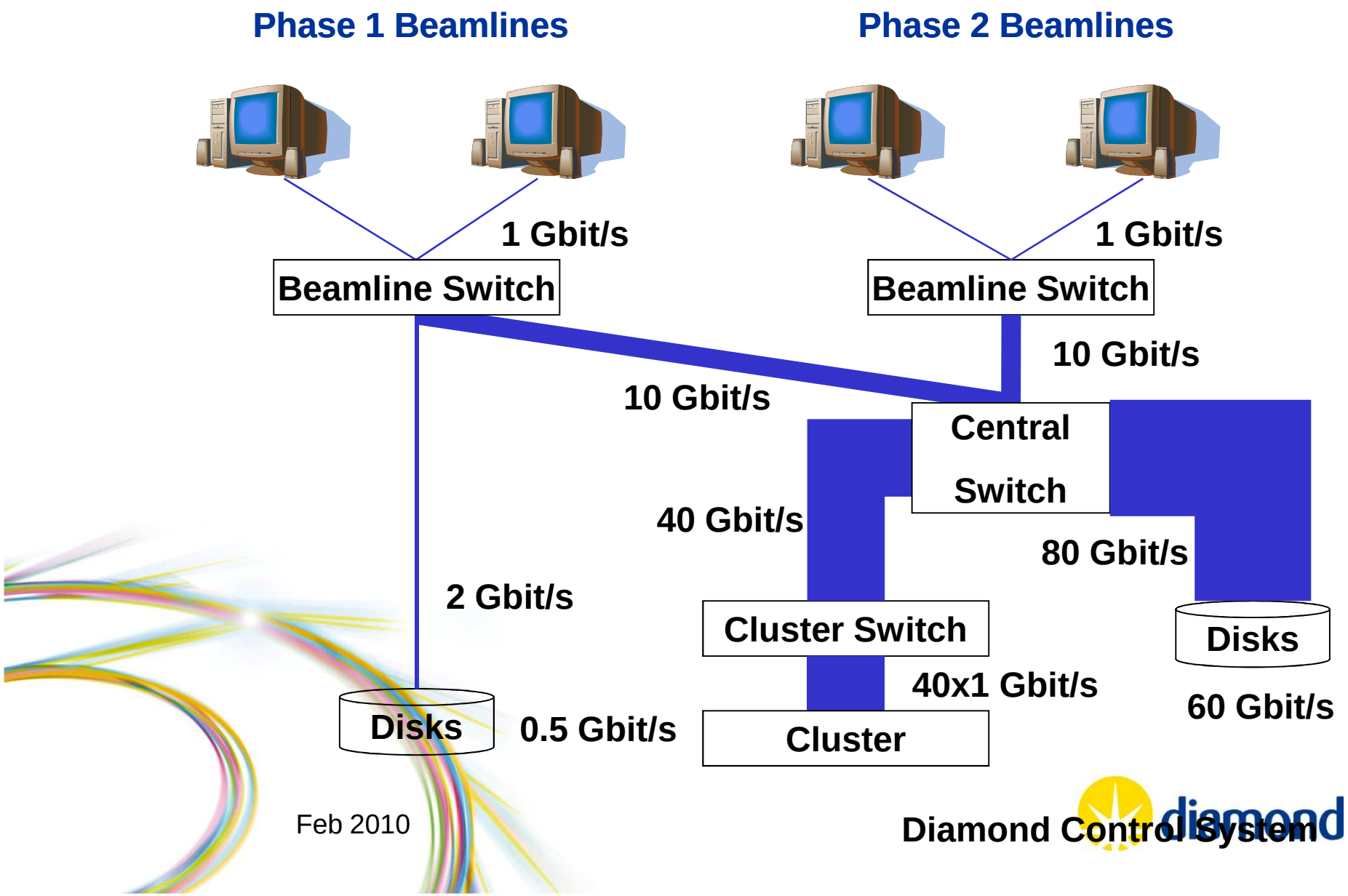
diamond

Network Layout



diamond

Network Layout



Network issues

- At high data rates packets get lost
- Standard TCP back off time is 200 ms, which is a lot of data at 10 Gbit/s
- See the problem mostly when 10 Gbit servers write to 1 Gbit clients.
- Other sites see similar issues
  - use your favorite search engine to look up “Data Center Ethernet” or “Data Center Bridging”

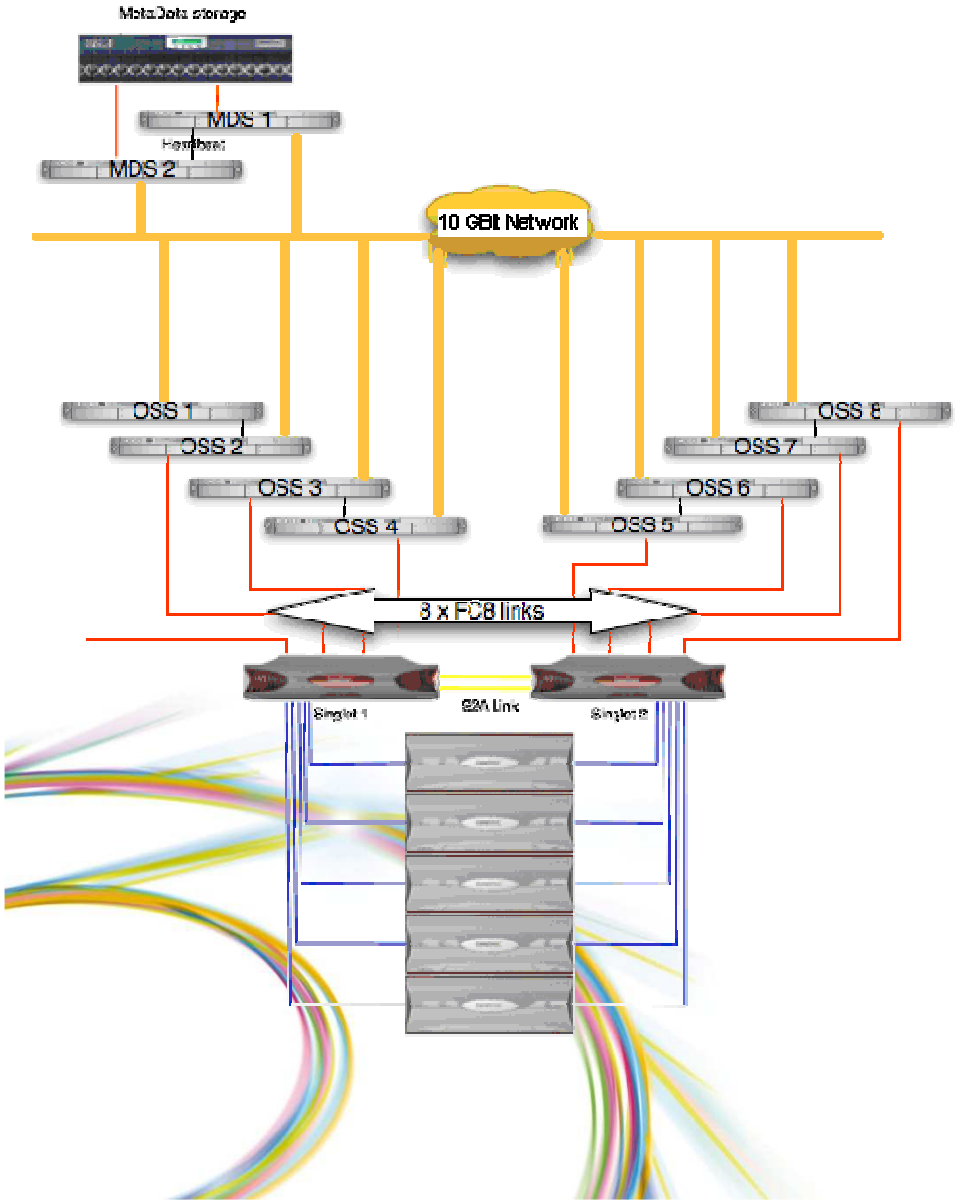
diamond

Storage – Requirements

- 200TBytes usable disk space
- 20 Mbytes/sec/Tbyte scalable aggregated throughput.
  - 4 GBytes/sec aggregated throughput for 200 Tbytes.
- 100 MBytes/sec transfer rate for individual 1 Gbit clients
- 400 MBytes/sec for individual 10 Gbit clients
- 8000 metadata operations/sec
- Highly resilient
- Support for Linux clients (RHEL4U6 and RHEL5 or later)
- POSIX with extended attributes and ACLs
- Groups for file access control based (> 256 groups/user)
- Ethernet Clients.
- Extendable by at least 200TB for future growth.

diamond

Storage - solution

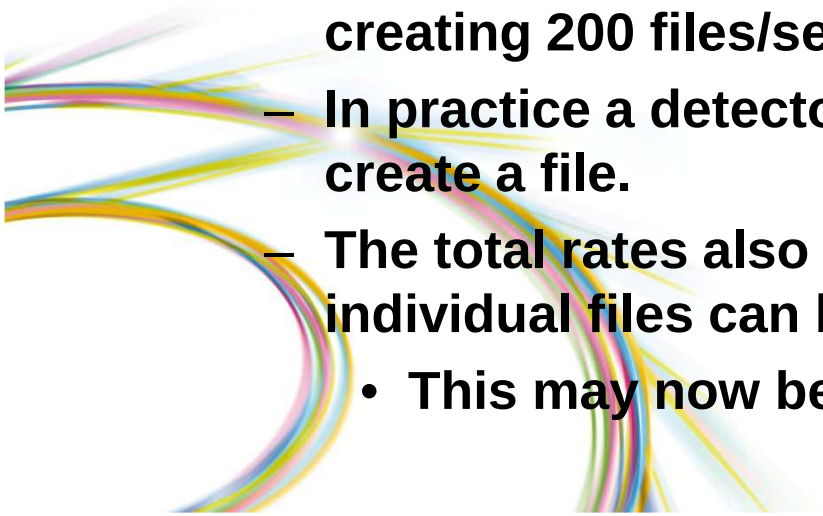


- Based on Data Direct Networks S2A9900 system
- Up to 5.7 GB/s throughput
- Fault-tolerant architecture
- Runs Lustre file system
  - Open source file system acquired by Sun in 2008, and now part of Oracle
  - Most popular file system in top 500 supercomputers

diamond

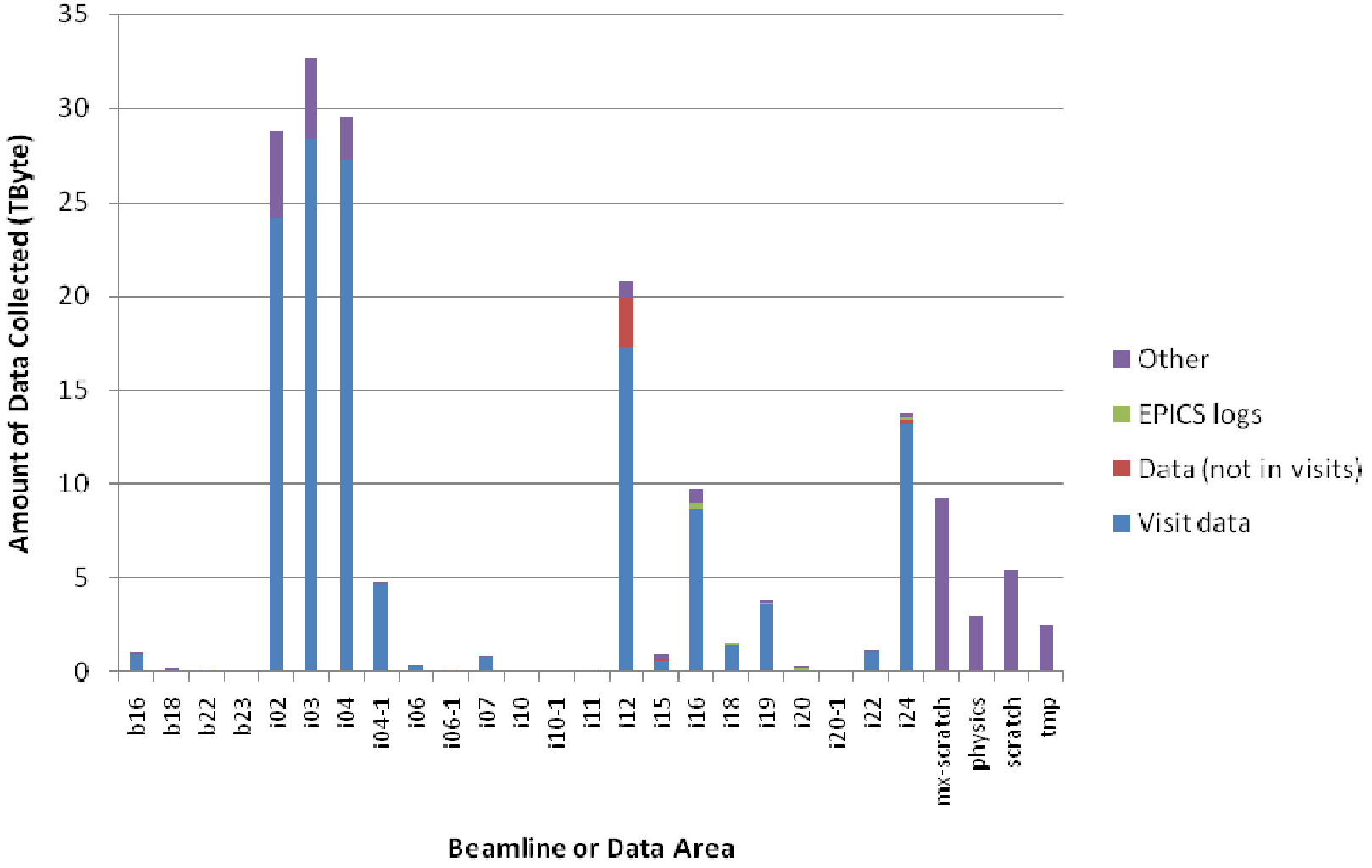
Storage - problems

- In initial tests:
  - Sometimes got near 4 Gbytes/sec writing
  - Only ever got about 3 Gbytes/sec reading
  - Only got 8000 metadata operations/sec for the specific case of sequential creates of zero length files.
  - Other metadata operations were slower (~2000/sec).
- In reality with real users:
  - Primarily seem to be limited by metadata rates
  - 8000 operations/sec specification based on 40 clients creating 200 files/sec.
  - In practice a detector takes 3-5 metadata operations to create a file.
  - The total rates also conceal that the latency for some individual files can be up to 15 seconds under load.
    - This may now be fixed...

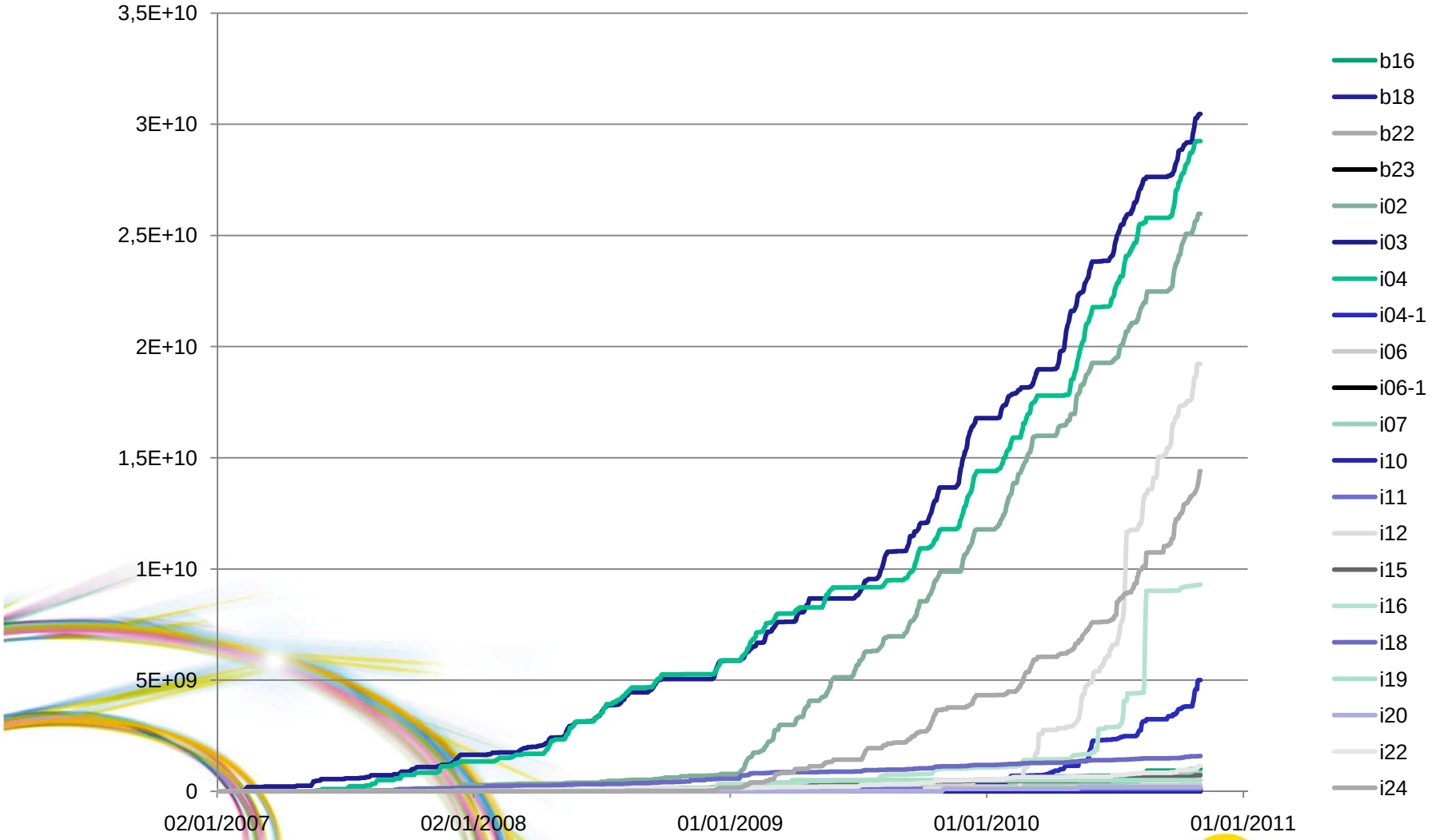


diamond

Storage status: Data sizes by beamline

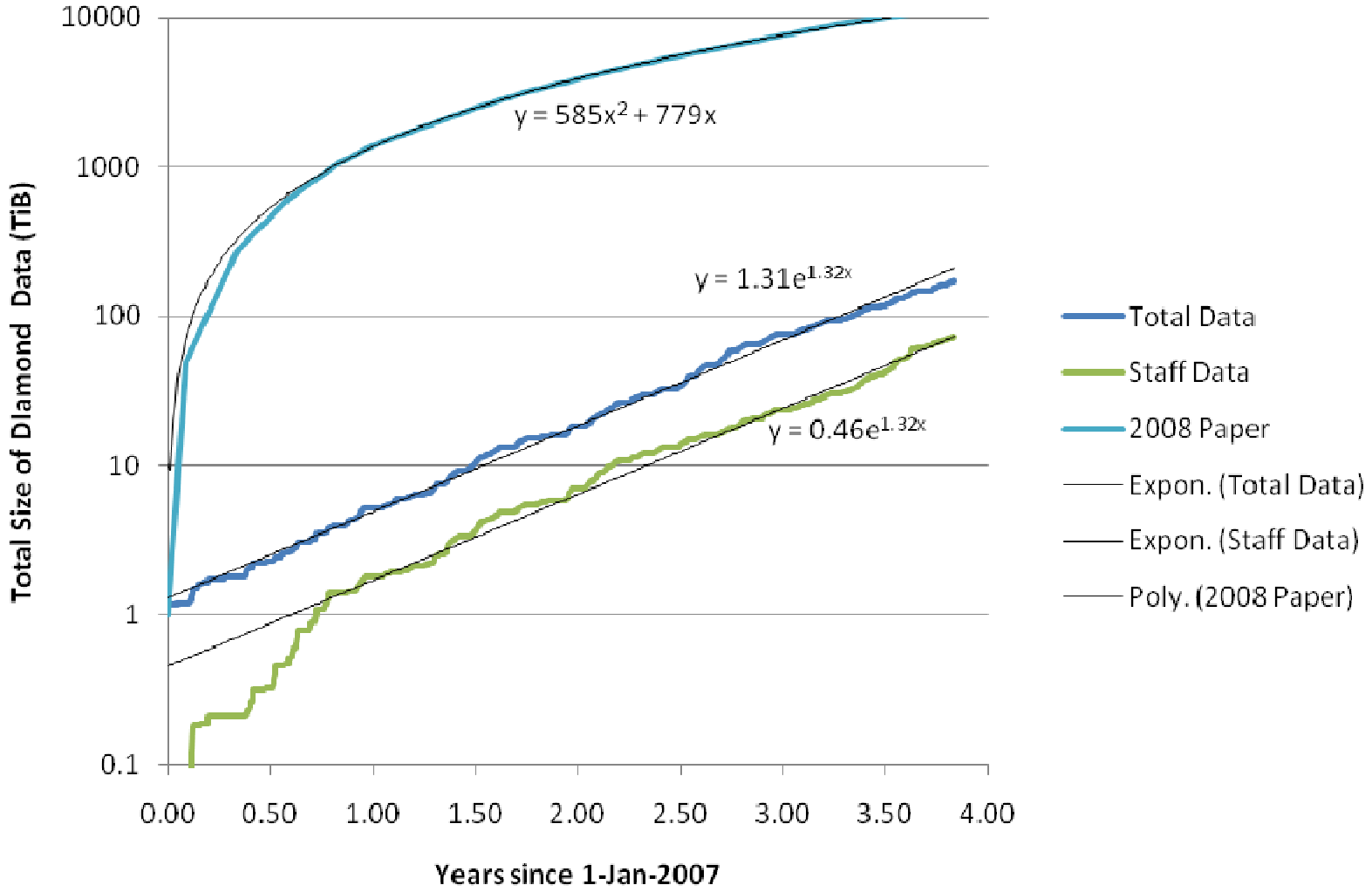


Data growth by beamline

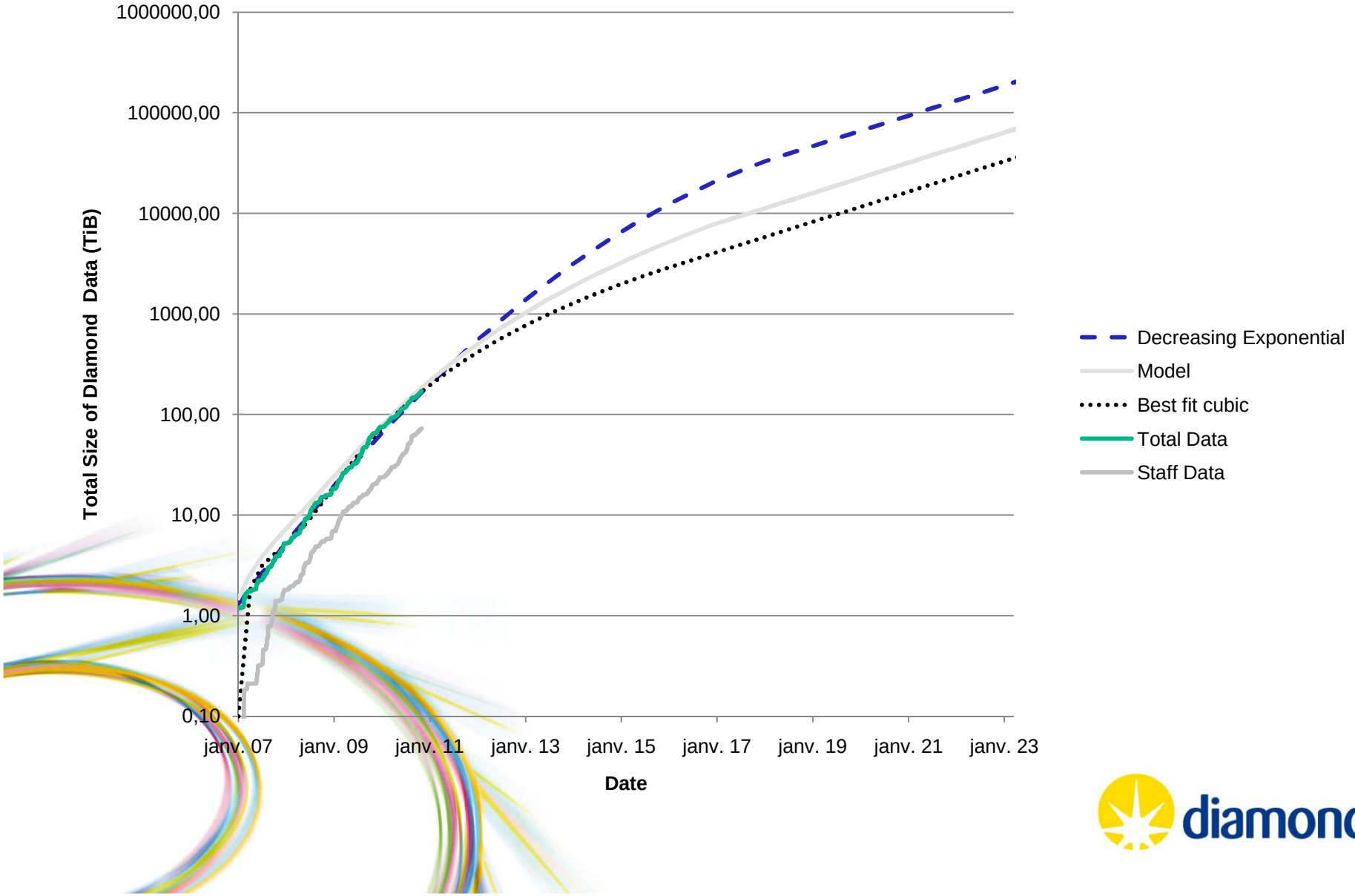


diamond

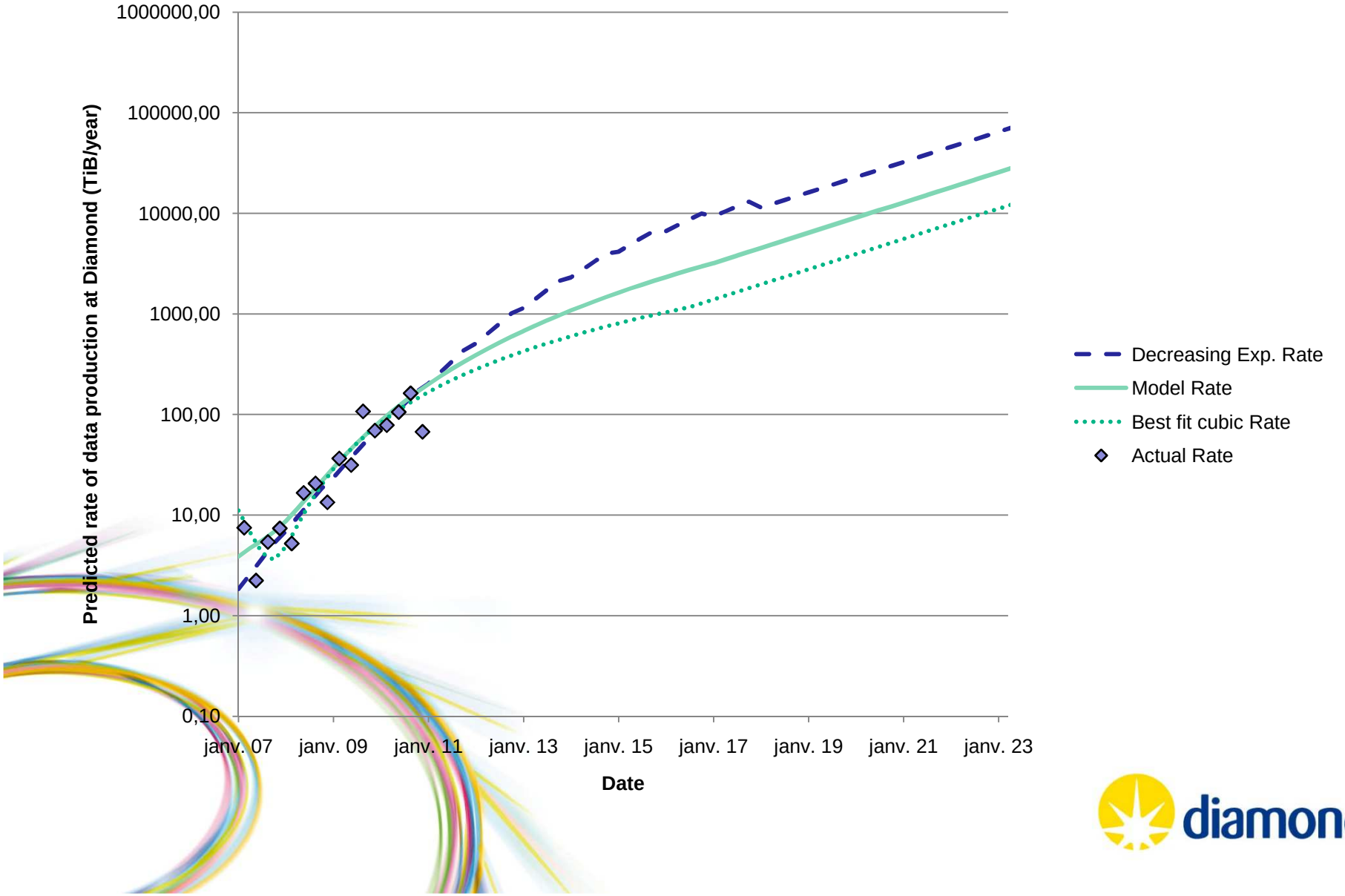
Total data growth



Where to next? – Data sizes



Where to next? - data rates



diamond

Where to next?

- Proposing to create a rolling programme for computing development based on Moore's Law and typical hardware lifetimes of 5 years
  - Replace most hardware every 5 years
  - Significant cluster and storage purchases every 2-3 years.
- A significant investment, but we must develop computing proactively in parallel with beamline data rates, rather than reactively because of them.

diamond

Conclusions

- **We have come from behind in computing but are gradually catching up.**
  - **Probably one of the most challenging synchrotrons with a large proportion of MX and an ambitious tomography group.**
  - **Must have long-term proactive approach to keep ahead of beamline requirements.**

diamond