Oh cr**p

... or when reality hits the CSC presentation material.

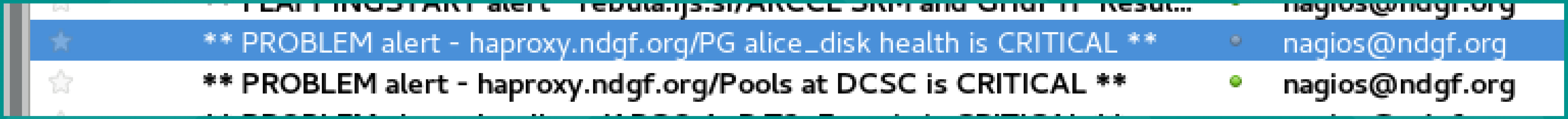
Ulf Tigerstedt
NDGF-T1/NeIC

NeIC
NORDIC E-INFRASTRUCTURE COLLABORATION

- NeIC/NDGF runs a WLCG Tier-1 for Atlas and Alice

- Funded by the nordic countries (except Iceland) but includes Slovenia (sort of)
- dCache used as storage software. Around 100 boxes, 9 Petabyte disk, ? Petabyte tape.
- Extremely distributed: storage hardware in Copenhagen, Ørestad, Linköping, Umeå, Oslo, Bergen, Espoo and Ljubljana

Monday afternoon:



** FLAPPINGSTART alert - rebula.ijs.si/ARCCE SRM and GridFTP Resul... nagios@ndgf.org

** PROBLEM alert - haproxy.ndgf.org/PG alice_disk health is CRITICAL ** nagios@ndgf.org

** PROBLEM alert - haproxy.ndgf.org/Pools at DCSC is CRITICAL ** nagios@ndgf.org

hpc_ku_dk_009, an 90TB “alice_disk” pool had failed with I/O errors, taking with it 70TB of Alice data.
The admin was reached on Tuesday morning, and his digging into the problem caused the machine to restart while at the same time the onboard out of band management locked up.

At this time, the filesystem turned up broken. All the xfs tools refused to touch it without risk, and the kernel refused to mount it without risk for more damage.

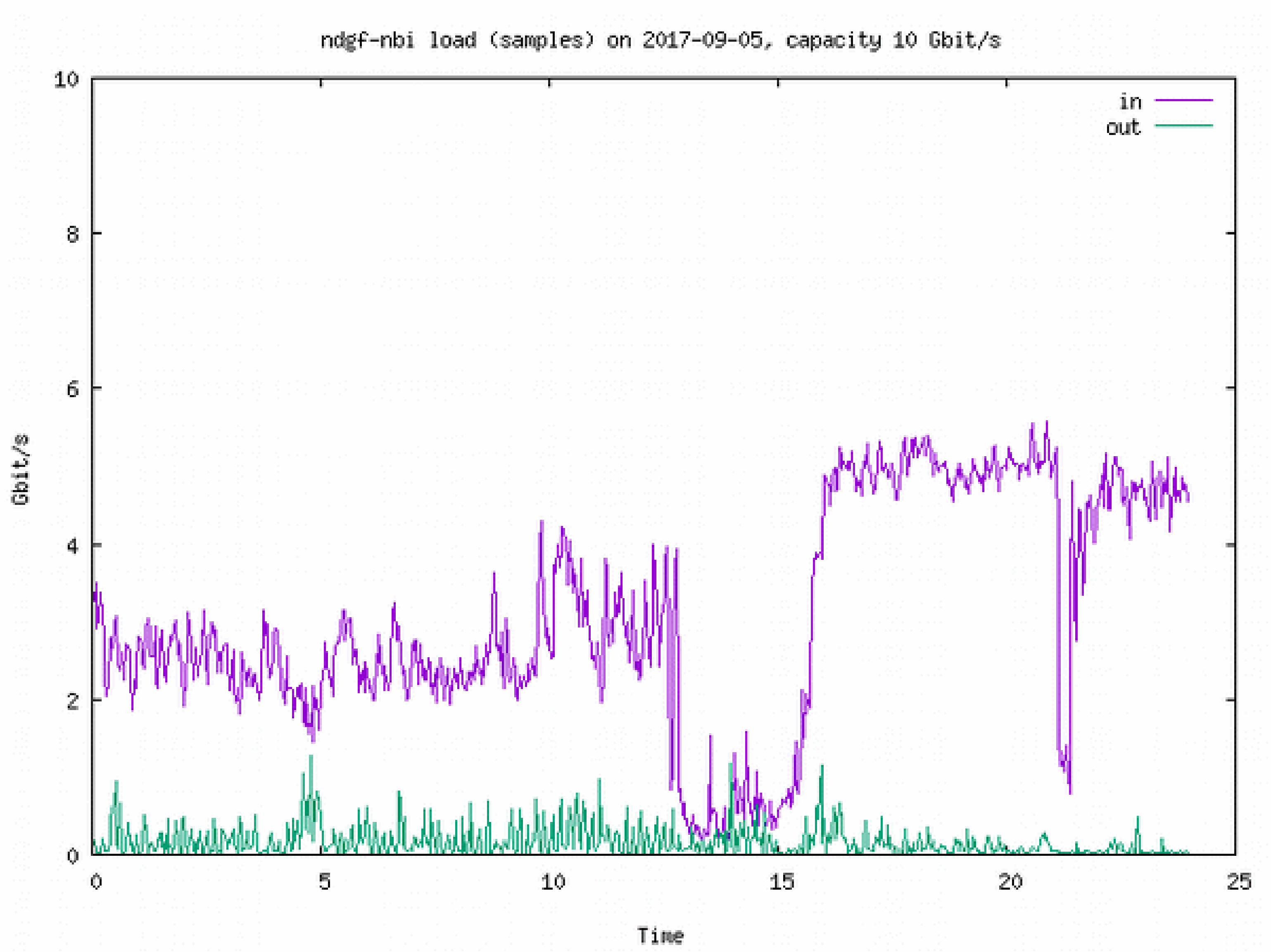
What now?

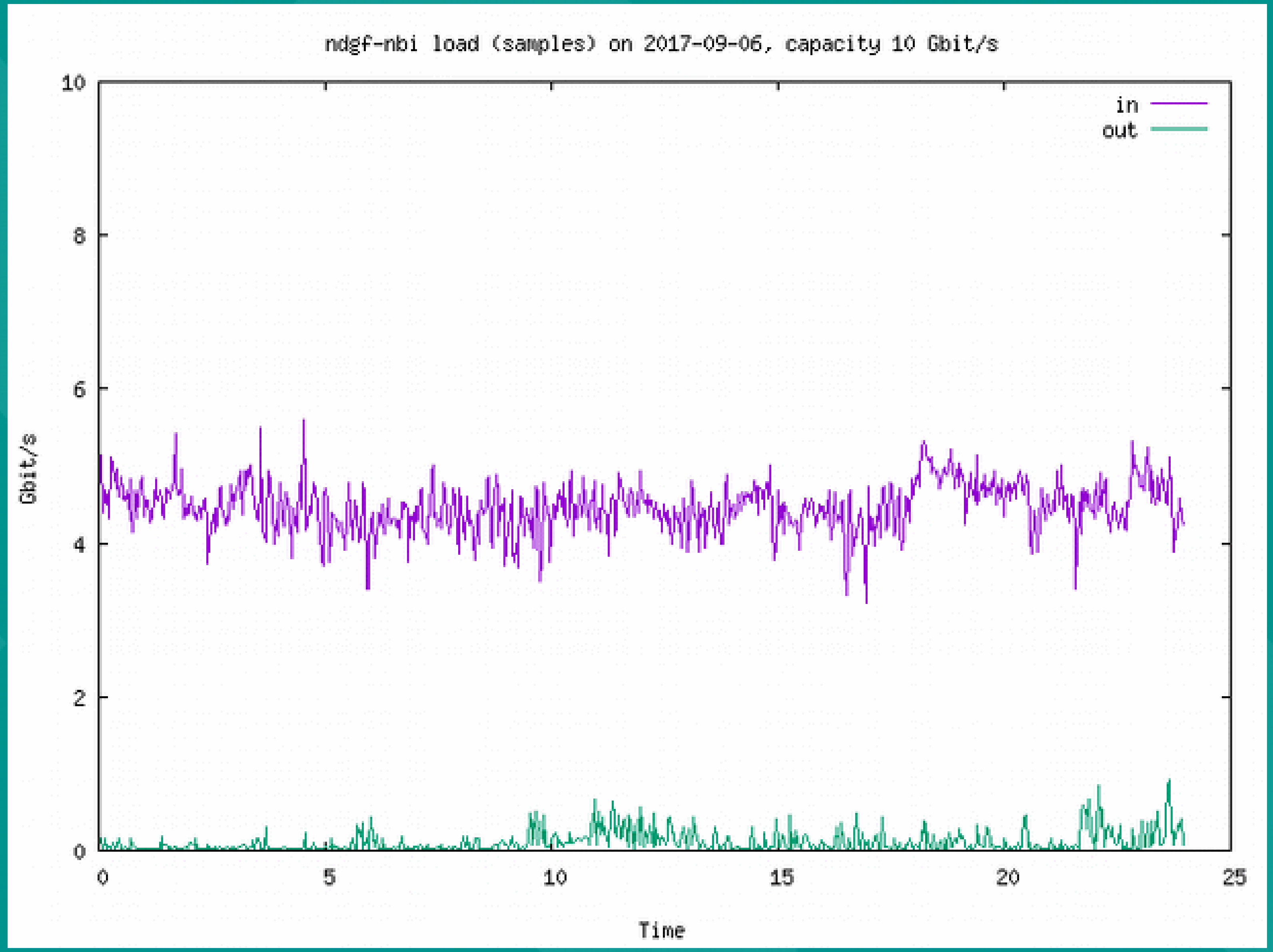
At this time, the filesystem turned up broken. All the xfs tools refused to touch it without risk, and the kernel refused to mount it without risk for more damage.

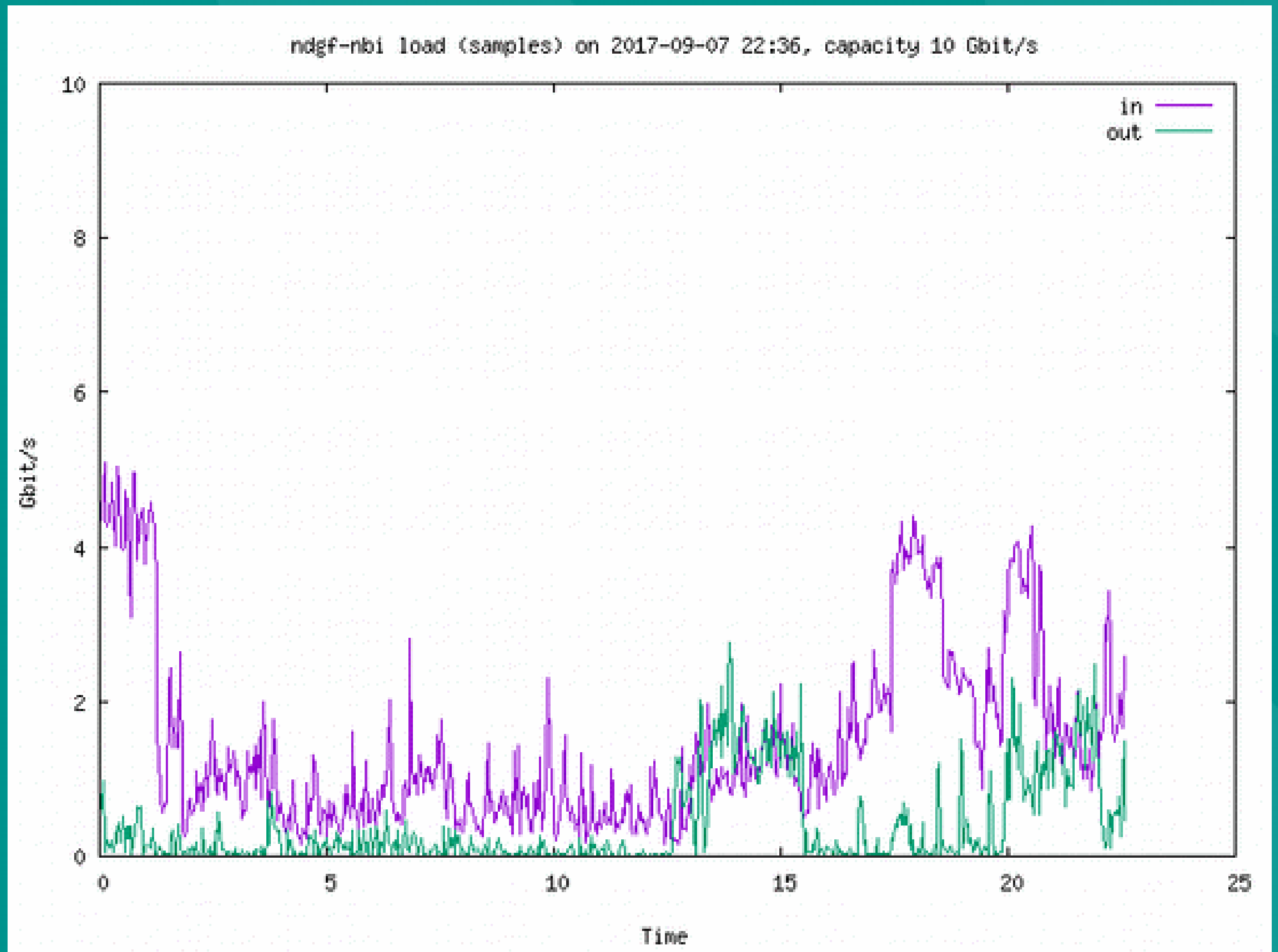
Start emptying anorther alice_disk pool on the same hardware. The data gets copied to two other machines in the same datacenter and to 3 new machines in Linköping.

450 MB/s, so 36~40 hours for 70TB.

2½ million files, so average filesize 30 MB







25 concurrent http transfers to keep everything moving nicely (TCP slow start, latency to Linköping, latency of namespace)

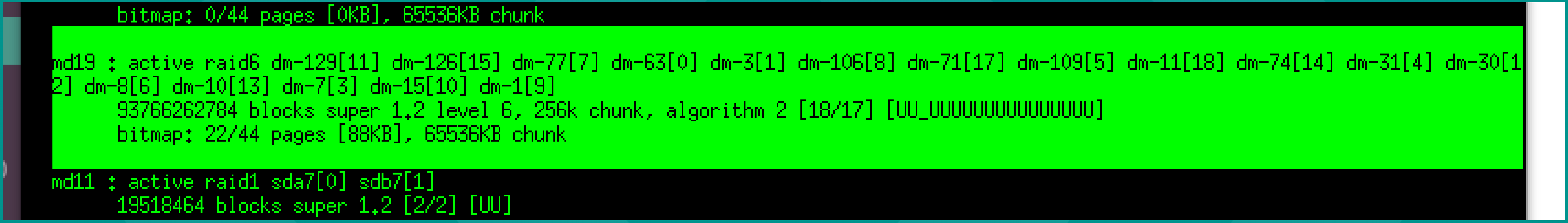
What's the hardware?

- Intel cpus, Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz
- 72 6TB disks in external disk shelf, some internal disks.
- DM multipath via 2 SAS controllers, 4 RAID6 sets of 18 disks each.

- Raidsets all looked ok after the reboot.

Or is it?

Wednesday morning::



```
        bitmap: 0/44 pages [0KB], 65536KB chunk

md19 : active raid6 dm-129[11] dm-126[15] dm-77[7] dm-63[0] dm-3[1] dm-106[8] dm-71[17] dm-109[5] dm-11[18] dm-74[14] dm-31[4] dm-30[1
2] dm-8[6] dm-10[13] dm-7[3] dm-15[10] dm-1[9]
        93766262784 blocks super 1.2 level 6, 256k chunk, algorithm 2 [18/17] [UU_UUUUUUUUUUUUUUU]
        bitmap: 22/44 pages [88KB], 65536KB chunk

md11 : active raid1 sda7[0] sdb7[1]
        19518464 blocks super 1.2 [2/2] [UU]
```

So what happened?

Corruption somewhere:
* one dying disk
* machine restarted by itself
* the onboard management also broke.
* filesystem

End result: 2.5 million/70TB of files gone.

Conclusion

- 18 disk RAID6 won't save you if there is a cascade of errors
- Doing anything on data of O(100T) takes ages
- Replicating between nodes migth not have saved this, since there seems to have been motherboard level corruption.