dCache hardware (+) layout

typical hardware usage for various scales

Jon Bakken FNAL

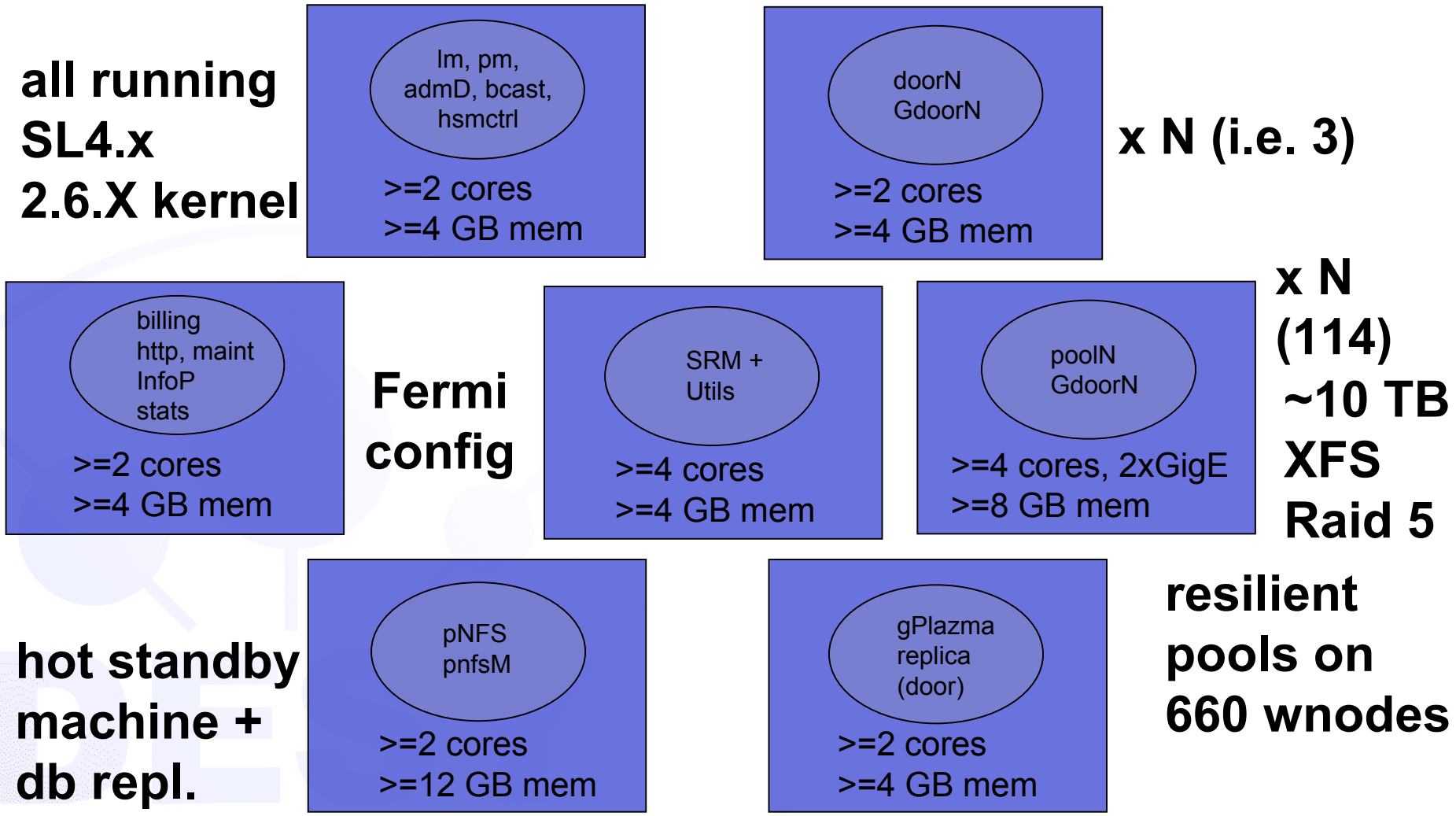
Martin Gasthuber DESY

runnable objects (threadgroups + processes) to spread out

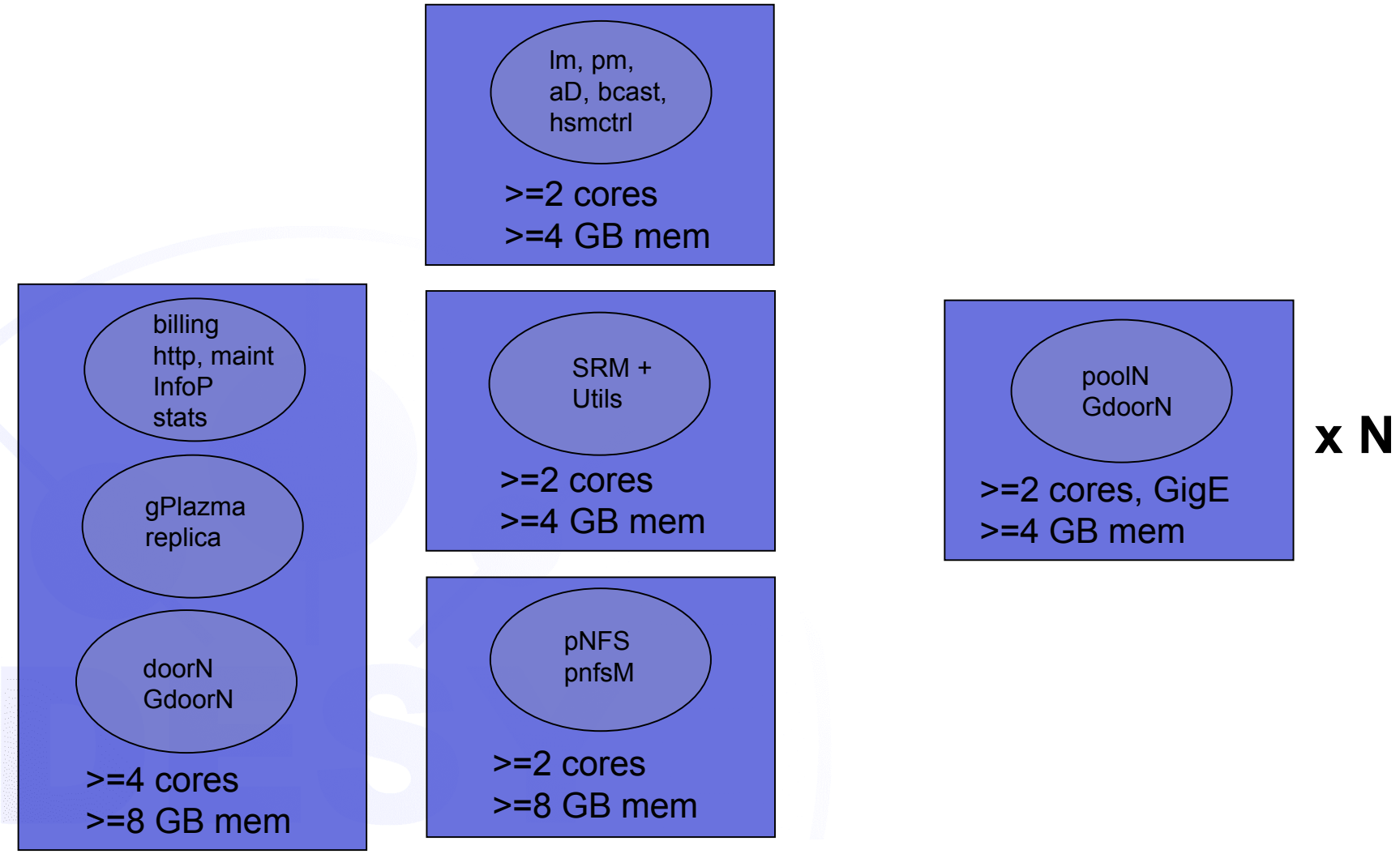
- PoolManager, lm 2
- Door (dCap, GFTP, adm, xrootd) 4+
- pNFS + pnfsManager 3+
- SRM, pinMan, spaceMan 3+
- billing, http, InfoProvider, gPlazma 3
- statistic, maint, hsmctrl, bcast 4
- pool, replica 2+

total: 21+

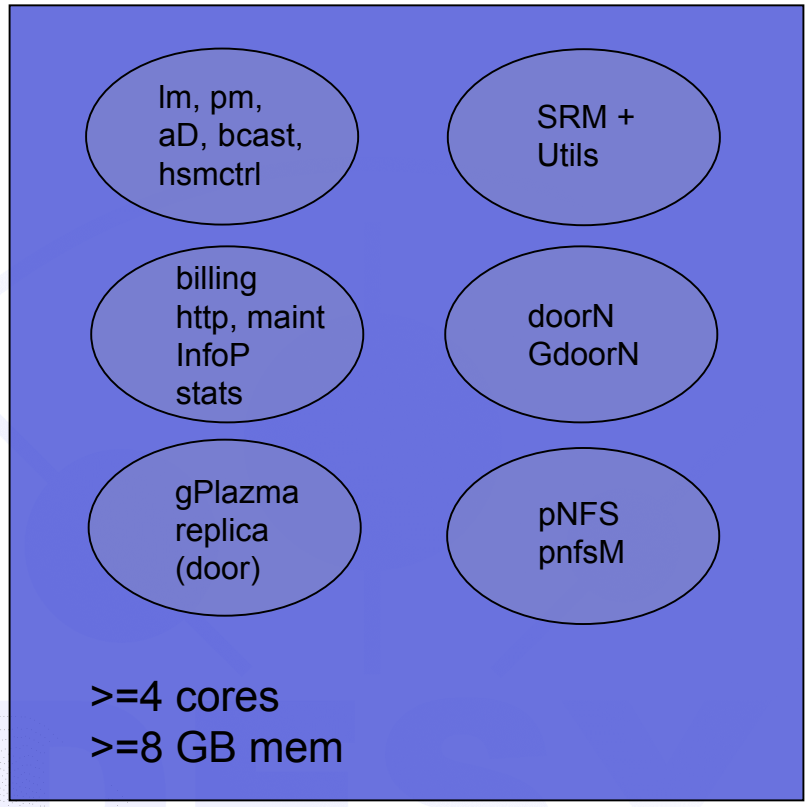
bubble view ... large installation

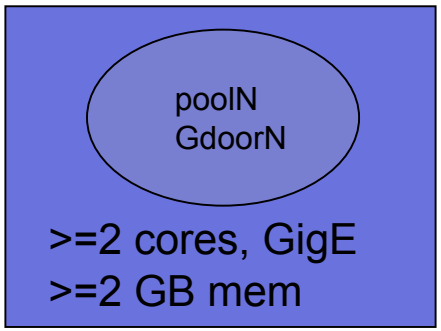


medium size ...


lm, pm,
aD, bcast,
hsmctrl
>=2 cores
>=4 GB mem
billing
http, maint
InfoP
stats
gPlazma
replica
doorN
GdoorN
>=4 cores
>=8 GB mem
SRM +
Utils
>=2 cores
>=4 GB mem
pNFS
pnfsM
>=2 cores
>=8 GB mem
poolN
GdoorN
>=2 cores, GigE
>=4 GB mem
x N

(very) small size



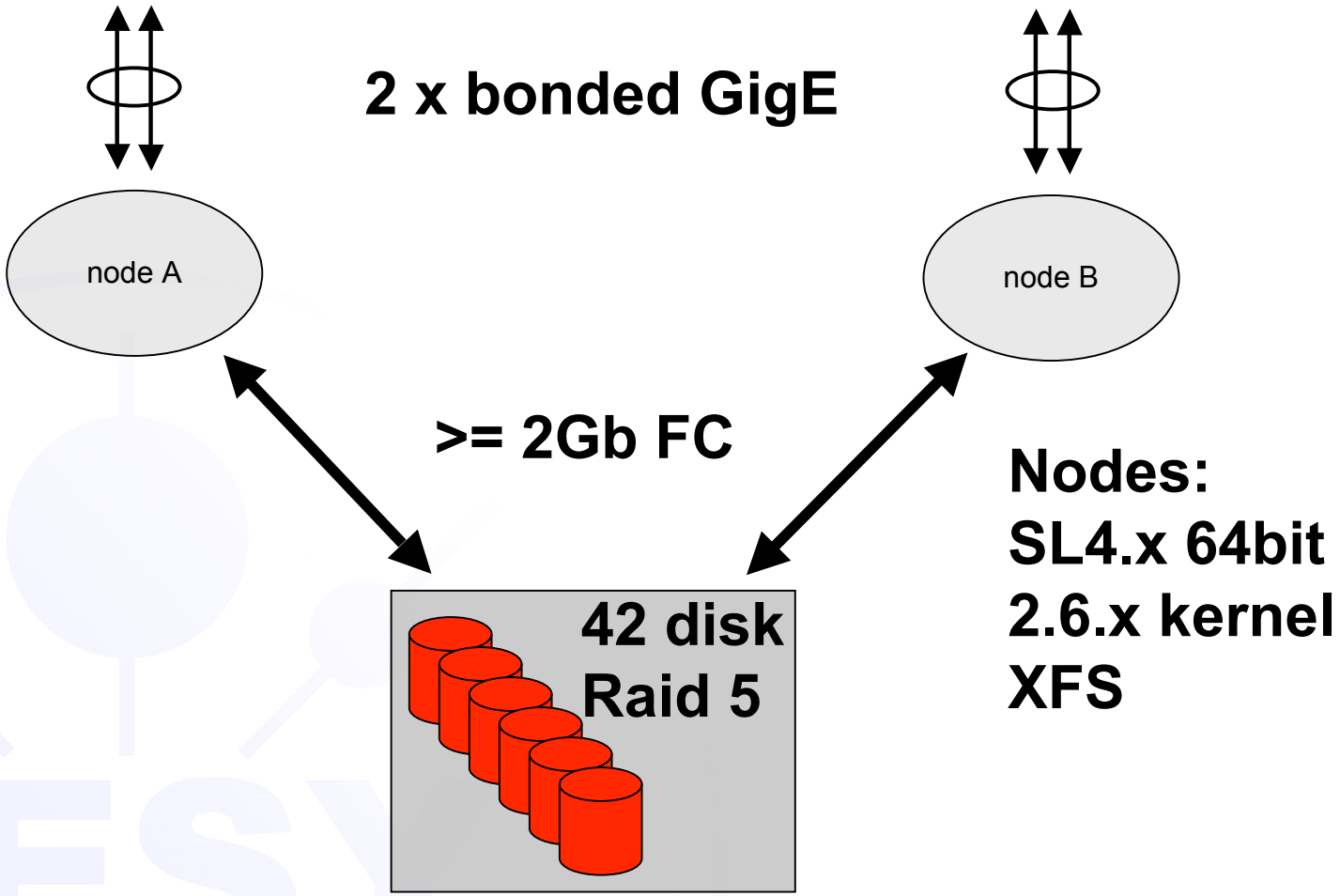


x N (1-4)

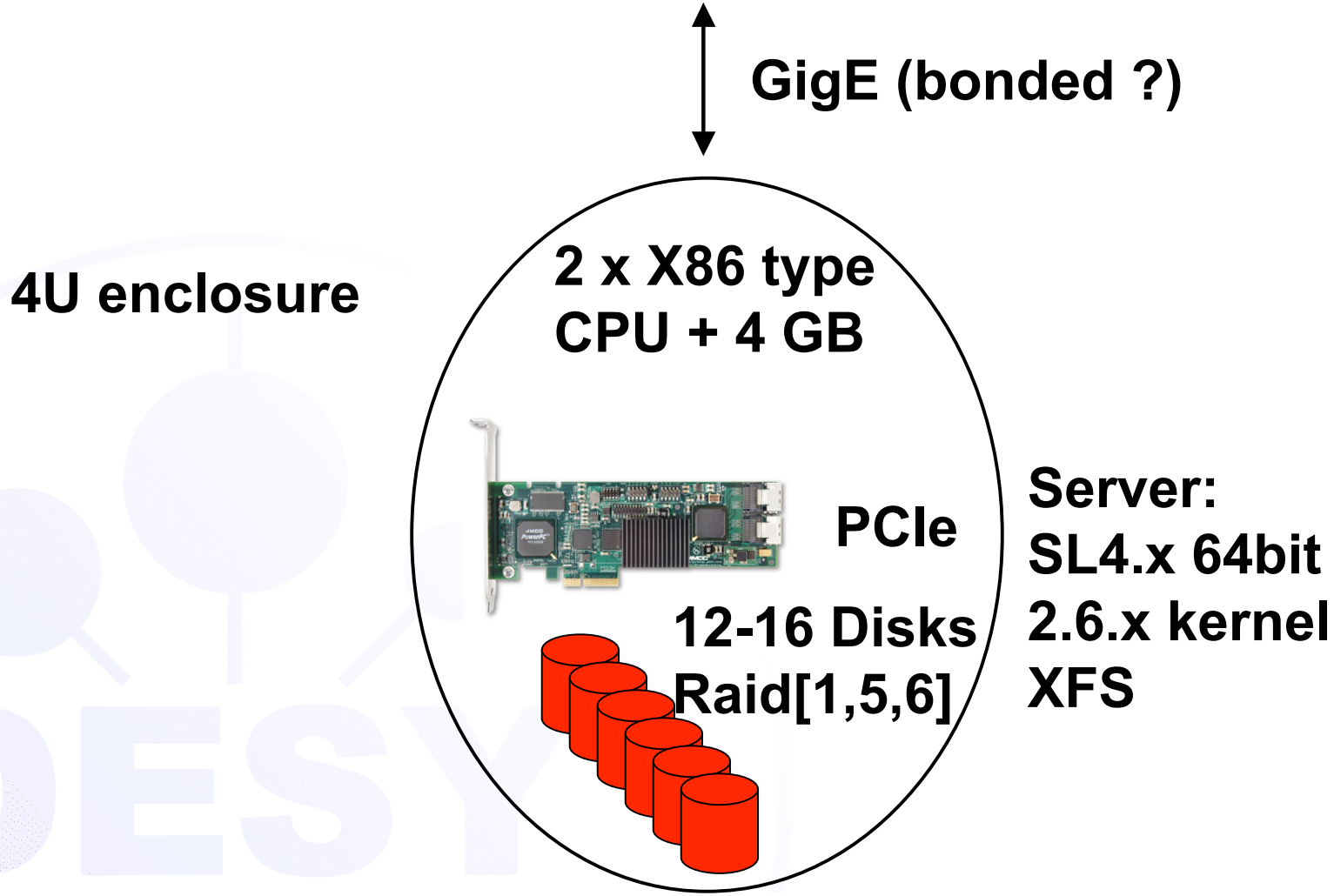
typical pool node types

- the ‘dual head external Raid’ system (Fermi)
  - *one leg* versions exists too
- Storage In a Box systems (cost effective)
  - classic
    - 3Ware, Areca PCI based multi-channel Raid N ctrl.
    - standard 2 cpu board - 4 GB memory - GigE
    - 12 to 16 SATA disks (250 - 500 GB)
  - in ‘initiation’
    - X4500 from Sun, 24 TB, ZFS (Solaris), 4xGigE

pool node config (Fermi typical)

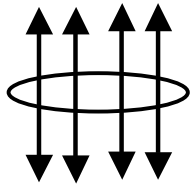


‘Storage in a Box’ … obvious !



‘Storage in a Box’ - new candidate !

SunFire X4500 (thumper)

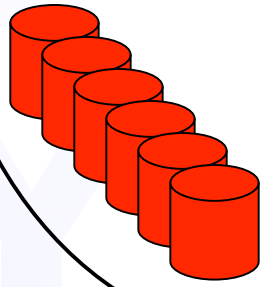


4 x GigE aggregation
(LACP based)

4U enclosure

4 Opteron
cores + 16 GB

6 SATA Contr.
8 channels each



48 Disks
mirror
single P
dual P

Server:
Solaris [10/11]
ZFS

how much - cores & mem

- looking at: *PoolManager, pNFS*
  - drives response times
  - heavy usage, critical part
- remaining cells/domains
  - 2 core + 4 GB mem are typical systems

observations - domains up ~1y

- PoolManager
  - 4 threads each 40% CPU
  - ~20 threads each 1-2% CPU
  - ~300 threads each ‘mainly nothing’ (~20min)
- statistics
  - 2 threads each 50% CPU
  - 5 threads each 1-2% CPU
- billing
  - 2 threads each 50% CPU
  - 5 threads each 1-2% CPU

4 core Opteron machine - 100% busy for 1 year !

observations - contd.

- pNFS + pnfsManager (USIII 6 CPU)
  - pnfsManager threads (5 threads - 40% CPU)
  - dbserver (gdbm) N x 1-2% CPU
  - pnfsd M x ~10% CPU

- 7 x pnfsd + 61 x dbserver + pnfsManager
  - needs >=4 cores and >=4 GB memory
  - more cores will help latency of 'ls' ;-)
    - not linearly !

observations - contd.

- high rate of ‘time slice exceeded’ - compared to ‘wait for resource’ (x3)
- too much ‘waiting for cpu’ state
- seq. thread (cpu) performance important

verifications

- will run same domain/cell configuration on a 16 core machine (same load)
- same for the pNFS/pnfsManager suite but different load environment.

- externalize DB storage (FC attached)
- more cores (>4) for PM, pNFS+

ref.

US-CMS T1 dCache (Fermi)
http://cmsdca.fnal.gov

http://www.dcache.org