# Evoluzione storage

**Andrei Maslennikov (CASPUR),**
**Vladimir Sapunenko (INFN/CNAF)**

**Maggio 2012 – Napoli**

# Disclaimer

- **Il tema «*Storage*» è un «bersaglio mobile» caratterizzato da un forte *feedback* positivo: lo sviluppo delle tecnologie di archiviazione e l'evoluzione delle apparecchiature capaci di produrre sempre più dati e in modo sempre più veloce si spingono ad oltranza.**

- **Questa presentazione non va intesa come un tentativo di sostituirsi a Google o Wikipedia o di coprire gli sviluppi a tutto campo. Abbiamo semplicemente cercato di fare un punto della situazione, menzionando alcuni fatti e alcune idee di potenziale interesse per questa comunità.**

# Sommario

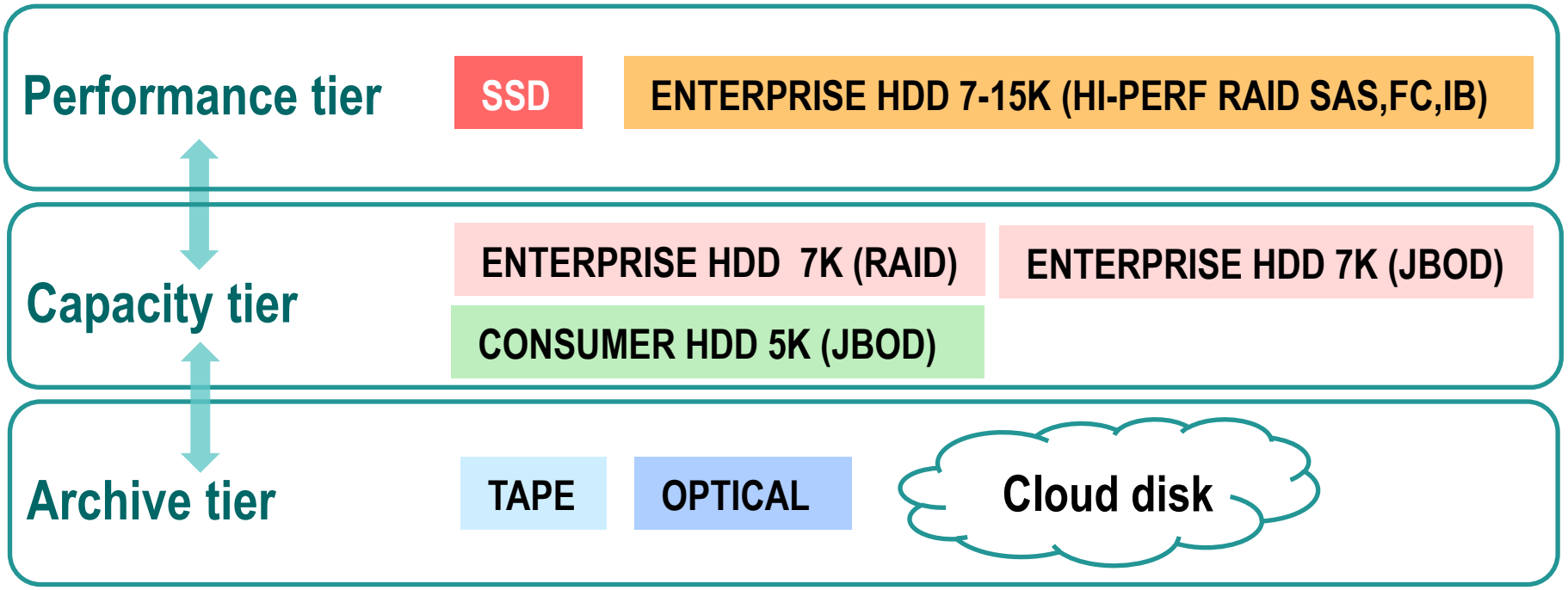
- Tendenze tecnologiche
- Building blocks: tape, HDD, SSD, cloud disk
- Stato di ben noti  file system distribuiti
- Qualche stima dei costi

# Tendenze tecnologiche

# Multi-tier: meno HSM, più backup, più cloud

- Il notevole calo del costo di 1 GB di HDD ha reso meno pronunciata la necessità di accesso automatico alla parte più lenta dell'archivio multi-tier locale (nastro). Il nastro viene piuttosto considerato come il media per backup asincrono e/o per archiviazione a lungo termine.

- Il cloud disk viene considerato come media di archiviazione per disaster recovery (business continuity).



# All'interno dei tier

**Performance tier**

- Meno futuro per i singoli sistemi RAID tradizionali;
- Soluzioni proprietarie scalabili, versatili, ricche di features
- Utilizzo diversificato dei dischi a stato solido;
- Utilizzo delle tecnologie ibride SSD/HDD;

**Capacity tier**

- Aumento di affidabilità e gestibilità dell'archivio attraverso l'intelligente distribuzione dei dati tra più nodi di servizio;
- Lato hardware: meno RAID, più JBOD, più SAS;

**Archive tier**

- Disaster recovery su cloud (D2D2C)

# Il tramonto del RAID classico

- **Con l'aumento della capacità del disco, i tempi necessari per ricostruire un volume RAID possono crescere a dismisura. Con i moduli da 2-4 TB i *rebuild* possono durare anche decine di ore, impattando sulle prestazioni e aumentando il rischio di perdita dei dati.**

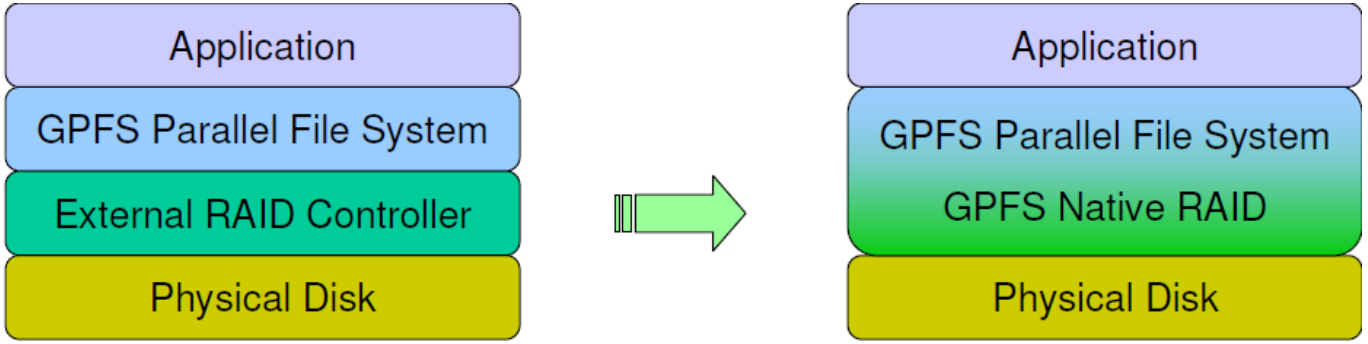
- **Il problema è dovuto al collo di bottiglia intrinseco all'architettura di un classico controller RAID: il processo di ricostruzione usa solo una parte degli IOPS disponibili per la lettura, mentre per la scrittura utilizza gli IOPS di un unico modulo.**

- **La soluzione al problema esiste: cambiando la politica di allocazione dei blocchi sui dischi fisici, si riesce a spalmare il traffico I/O uniformemente su tutti i moduli e sfruttare tutti gli IOPS disponibili.**

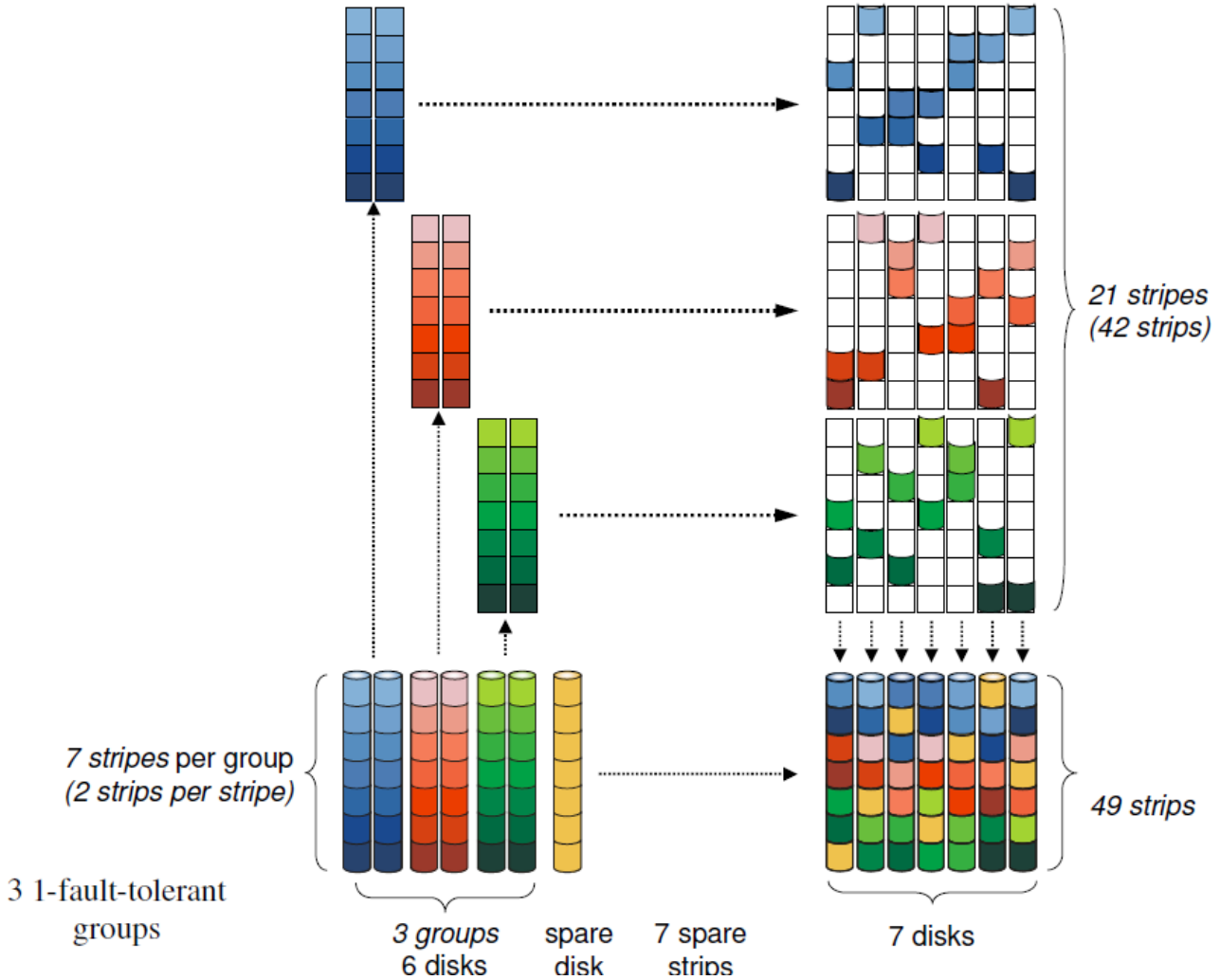
# Un esempio: declustered RAID in GPFS

- **La IBM è stata una delle prime ad implementare la nuova strategia all'interno del suo famoso prodotto, GPFS. Sopranominata «Declustered RAID», la soluzione è realizzata in software e si poggia sui JBOD invece che sui classici volumi RAID hardware.**

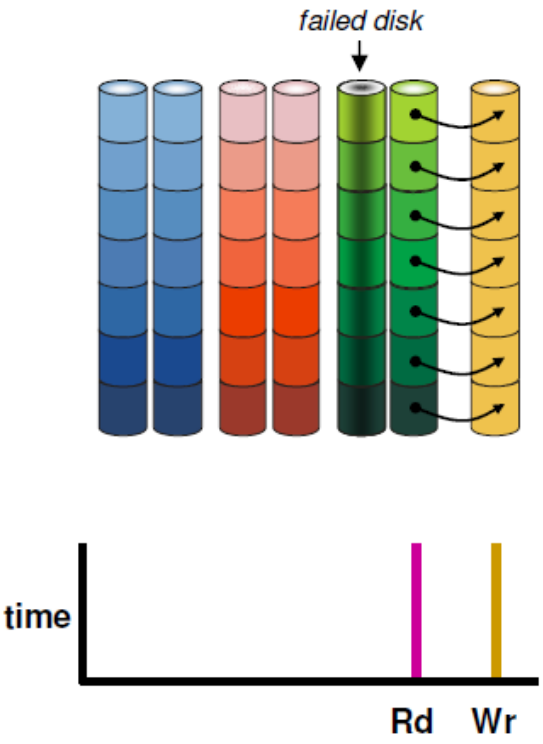
- **NB: la soluzione è praticabile solo sotto l'AIX. Gli RAID LUN esistenti possono essere convertiti nei JBOD.**



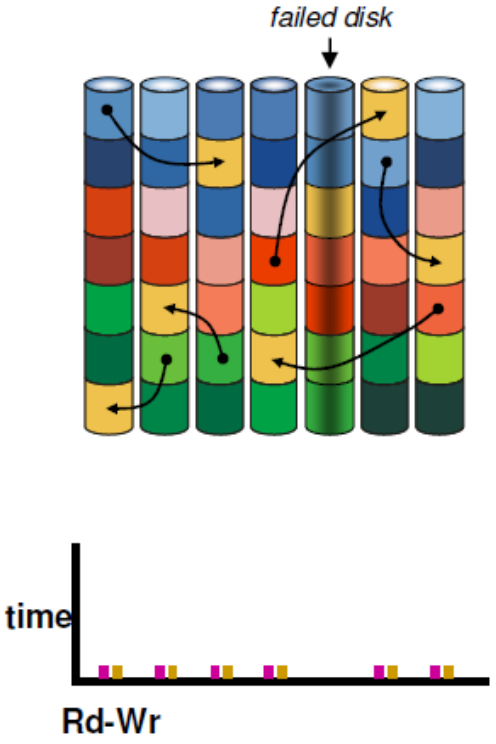
# RAID 1 classico vs declustered



# Rebuild di un RAID1 (classico vs declustered)

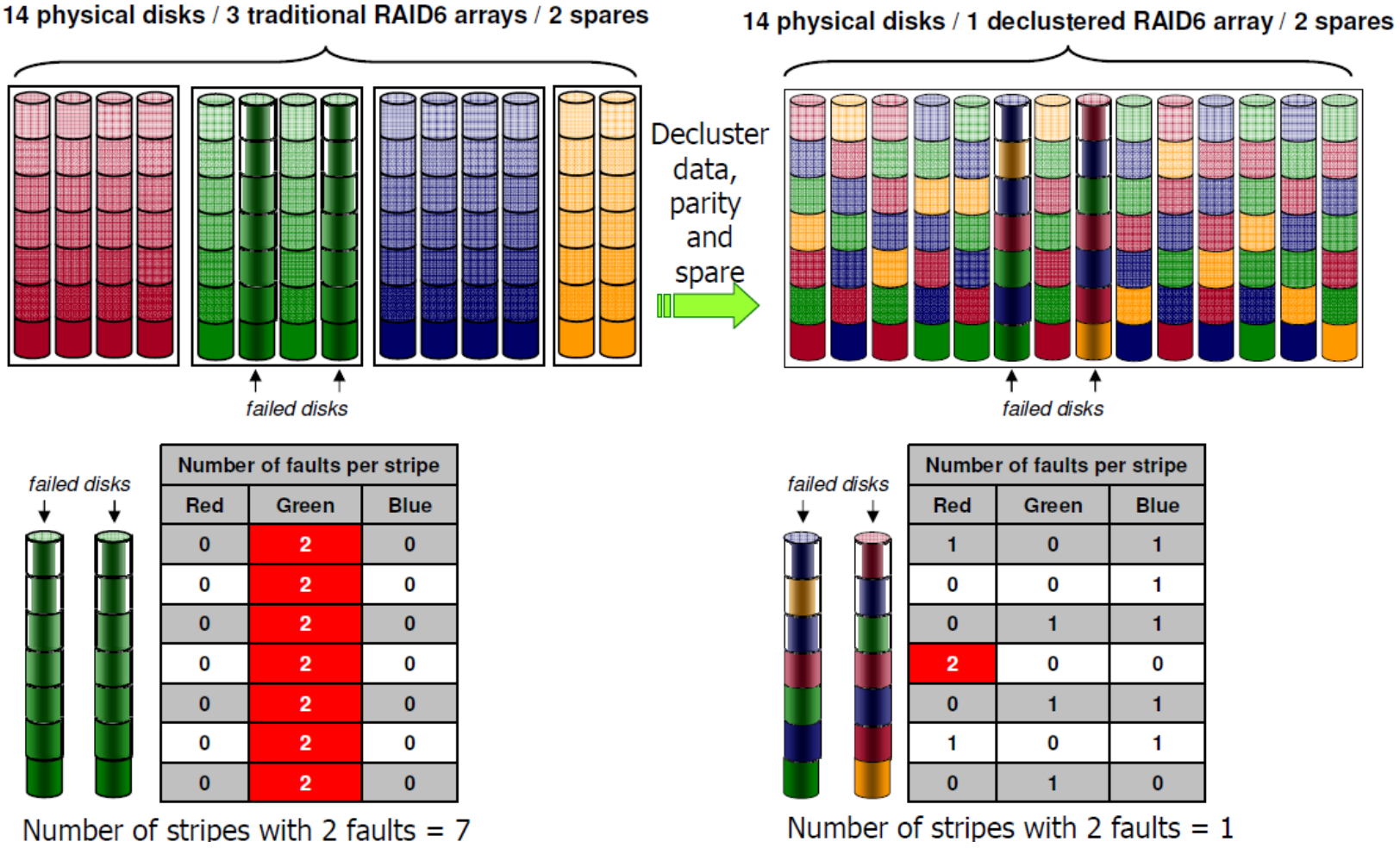



Rebuild activity confined to just a few disks – slow rebuild, disrupts user programs




Rebuild activity spread across many disks, faster rebuild or less disruption to user programs

# RAID6 (classico vs declustered)



# Declusterizzazione in rete

- **Oggi sono diventate diffuse le soluzioni di storage scale-out che estendono il principio di declusterizzazione oltre il confine di un nodo di servizio e, con poche eccezioni, non fanno uso di hardware RAID. (Concettualmente il file viene visto come un «cluster» di più «chunk» di dati).**

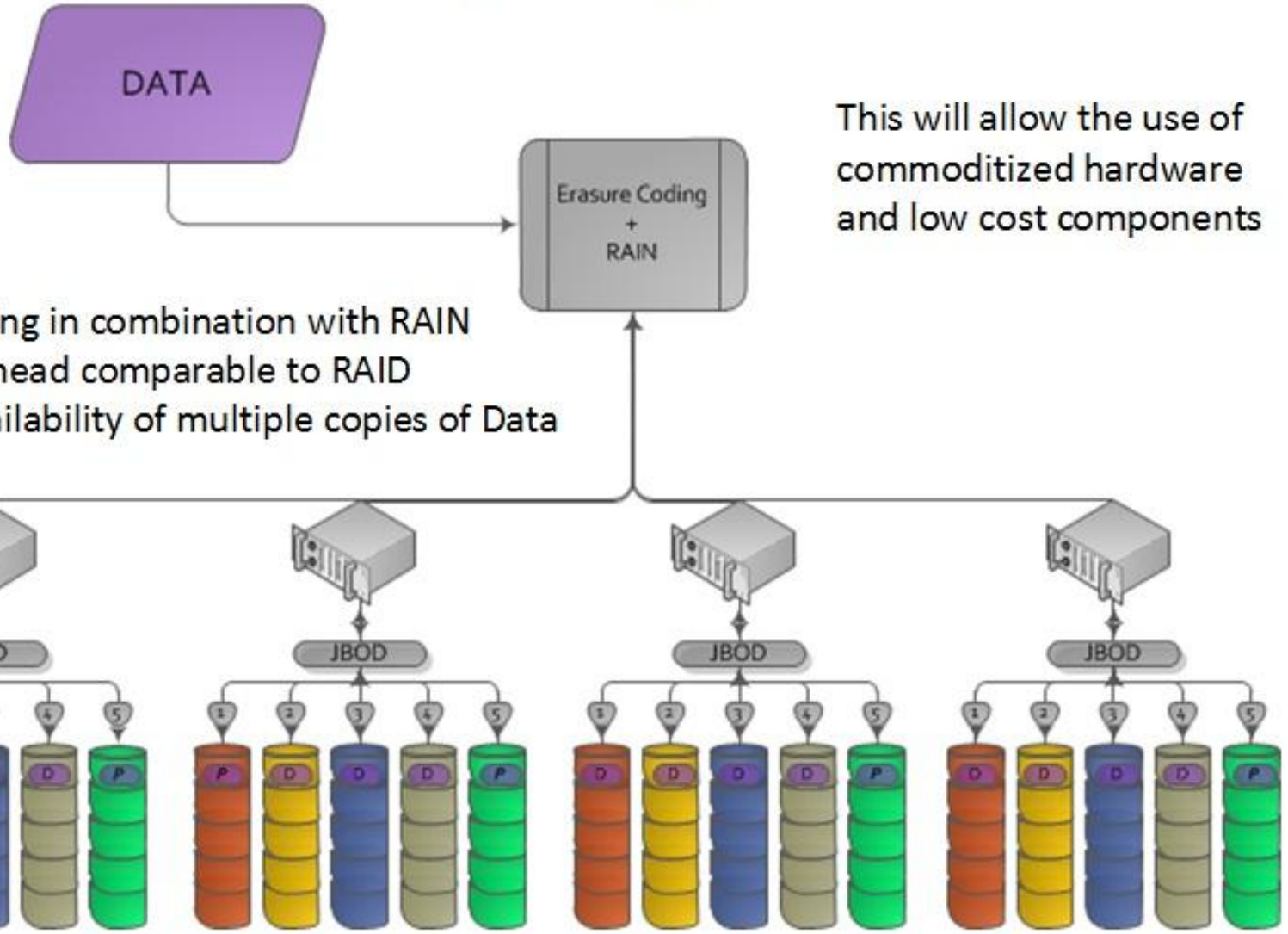
- **Declusterizzando con l'eccesso le informazioni archiviate (con l'aiuto di tecniche di replicazione o erasure coding) si riesce a garantire la disponibilità dei dati anche nei casi di indisponibilità dei nodi di servizio.**

- **Buzzwords: network parity (coniato da Panasas), RAIN (Redundant Array of Independent /Inexpensive Nodes).**

- **Alcune soluzioni che fanno uso dei RAIN: Google FS, GlusterFS, AFS/OSD, CERN EOS, IBM XIV, Panasas, DDN WOS, EMC Isilon, Permabit ..**

# RAIN

## Redundant Array of Inexpensive Nodes



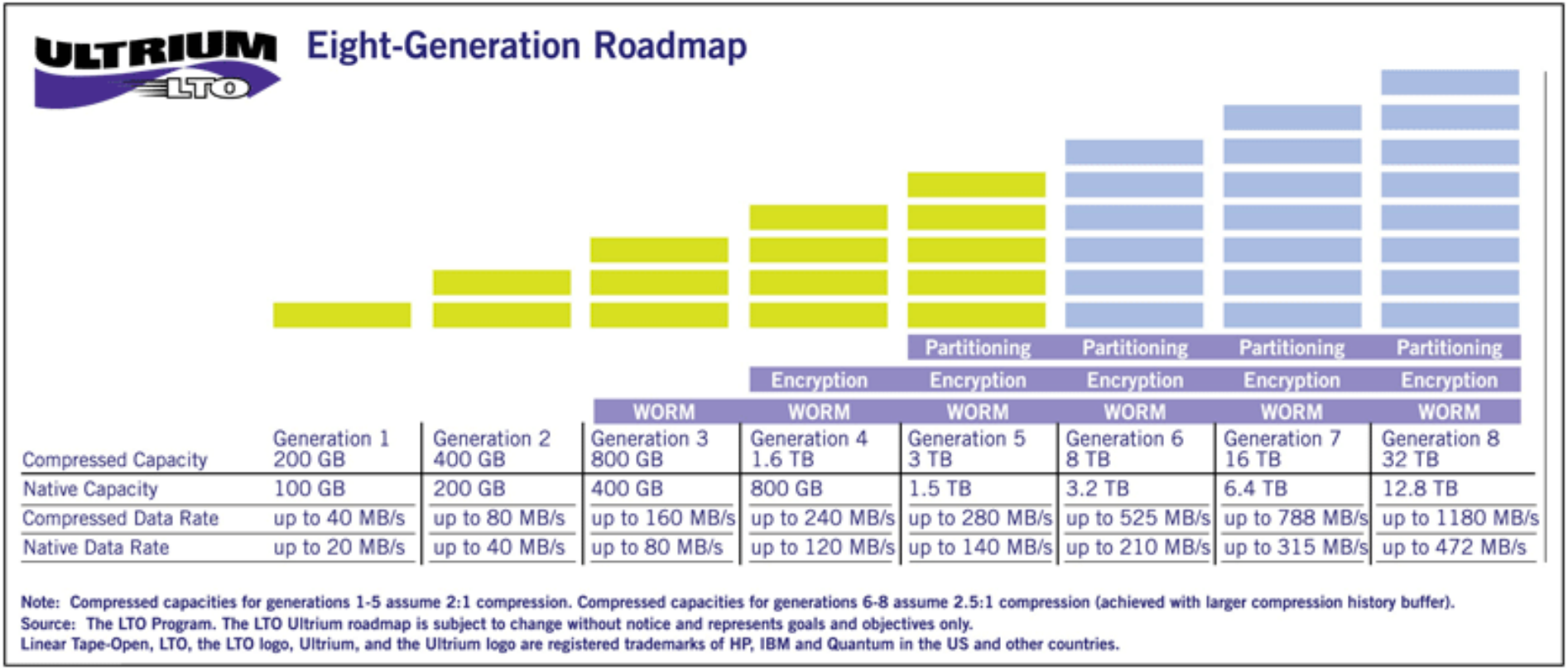

This will allow the use of commoditized hardware and low cost components

Erasure Coding in combination with RAIN allows overhead comparable to RAID with the availability of multiple copies of Data

# Storage building blocks

# Tape 2012

- LTO sta dominando il mercato con il market share di 90%. LTO-6 sarà commercializzato a partire dal Q3 2012.
- Introdotto con LTO-5: supporto per Linear Tape File System (LTFS). Cartucce formattate con LTFS e montate si presentano come cartelle direct access con metadata caching. Open source.




| Compressed Capacity | Generation 1<br>200 GB | Generation 2<br>400 GB | Generation 3<br>800 GB | Generation 4<br>1.6 TB | Generation 5<br>3 TB | Generation 6<br>8 TB | Generation 7<br>16 TB | Generation 8<br>32 TB |
|---|---|---|---|---|---|---|---|---|
| Native Capacity | 100 GB | 200 GB | 400 GB | 800 GB | 1.5 TB | 3.2 TB | 6.4 TB | 12.8 TB |
| Compressed Data Rate | up to 40 MB/s | up to 80 MB/s | up to 160 MB/s | up to 240 MB/s | up to 280 MB/s | up to 525 MB/s | up to 788 MB/s | up to 1180 MB/s |
| Native Data Rate | up to 20 MB/s | up to 40 MB/s | up to 80 MB/s | up to 120 MB/s | up to 140 MB/s | up to 210 MB/s | up to 315 MB/s | up to 472 MB/s |

Note: Compressed capacities for generations 1-5 assume 2:1 compression. Compressed capacities for generations 6-8 assume 2.5:1 compression (achieved with larger compression history buffer).
Source: The LTO Program. The LTO Ultrium roadmap is subject to change without notice and represents goals and objectives only.

# Il ruolo dei nastri cambia, ma sono indispensabili

- I nastri magnetici non sono più visti come la parte attiva dell'archivio. La ragione principale di questo mutamento di ruolo è l'enorme calo di costo di disco nearline.

- Però, il nastro ha i suoi punti di forza: con longevità di circa 50 anni, sono ottimi per «deep archiving». Inoltre, il consumo energetico di una libreria a nastri è di gran lunga inferiore di quello di un sistema a dischi di simile capacità.

- I vendor dichiarano un fattore anche di 6.5 tra il TCO di un archivio basato su nastri e di un altro di capacità analoga basato sui HDD.

Assumendo il costo di 15 centesimi per 1 KWh si ottiene almeno il fattore 2:

- 3 anni di TCO di una libreria SpectraLogic T680 piena con 4 lettori LTO-6 costerebbero circa 166 KE (2.8 PB compressi in 42 RU, 600 W);

- 3 anni di TCO di 867 HDD nearline da 3TB in dense JBOD costerebbero circa 320 KE (2.8 PB raw in 56 RU, 17000 W, senza replicazione o erasure coding);

# HDD 2012

- Dischi fissi classici SATA/SAS di qualità consumer (C, 5400 RPM) ed enterprise (E, >= 7200 RPM) da 3.5/2.5 pollici. Principali produttori: WD/Hitachi e Seagate/Samsung.

| RPM | Max TB | IOPS | Costo per terabyte 2012, Euro (*) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | 0.45 T | 0.6 T | 0.9 T | 1 T | 2 T | 3 T | 4 T |
| C 5400 | 3 | 60 | | | | | | 50 | |
| E 7200 | 4 | 100 | | | | 100 | 107 | 137 | 137 |
| E 10000 | 0.9 | 150 | | 480 | 700 | | | | |
| E 15000 | 0.6 | 200 | 560 | 736 | | | | | |

- Seagate è stato il primo a raggiungere la densità di 1 terabit / pollice quad. (marzo 2012, Heat Assisted Magnetic Recording). Next drive capacity: 6TB. Potenziale: fino a 50-60TB per drive tra 10+ anni.

(*) Origine prezzi citati: E4 e altri

# SSD 2012

- **Non-consumer SSD: EMLC, SLC, DRAM. Lunga lista di produttori: Intel, Samsung, Fusion-IO, OCZ, STEC ,TMS, NextIO, Violin etc.**

**Prestazioni massimali (aprile 2012):**
**- Rackmount 4U: KOVE Xpress Disk (RAM): 28 GB/sec(IB), 11.7M IOPS**
**- PCI Express : Fusion-io ioDrive Octal (Flash): 6.2 GB/sec, 1M IOPS**
**- 3.5 poll. RAM : STEC Zeus RAM : 100K/100K IOPS**
**- 3.5 poll. EMLC: STEC Zeus IOPS: 550/300 MB/sec, 80K/40K IOPS**

**Alcuni esempi di SSD:**

| Modello | GB | IOPS 4K (R/W), MB/s | TBW,TB | E/TB |
|---|---|---|---|---|
| GENERIC MLC CONSUMER SATA | 400 | 80K/70K, up to 500 MB/s | 70 | 1200 |
| SAMSUNG SM825 (EMLC, SATA) | 400 | 43K/11K, 200 MB/s | 7000 | 2000 |
| INTEL 910 (MLC-HET, PCIE) | 800 | 180K/75K, 750 MB/s | 14000 | 3700 |
| STEC ZEUSIOPS (EMLC,SAS) | 800 | 115K/70K, up to 500 MB/s | 33000 | 4800 |
| FUSION IODRIVE 2 (SLC, PCIE) | 1200 | 175K/235K (norm), 3GB/s | N/A | 20+K |
| STEC ZEUSRAM (DRAM, SAS) | 8 | 100K/100K, 900MB/s | +++ | 240K |

# Progressi SSD

- I flash memory controller diventano sempre più intelligenti:
  - Miglioramento di algoritmi di wear levelling;
  - Utilizzo dei blocchi più piccoli (512 B);
  - Write multiplication più basso (Intel: 1.1);
  - Built-in auto healing;

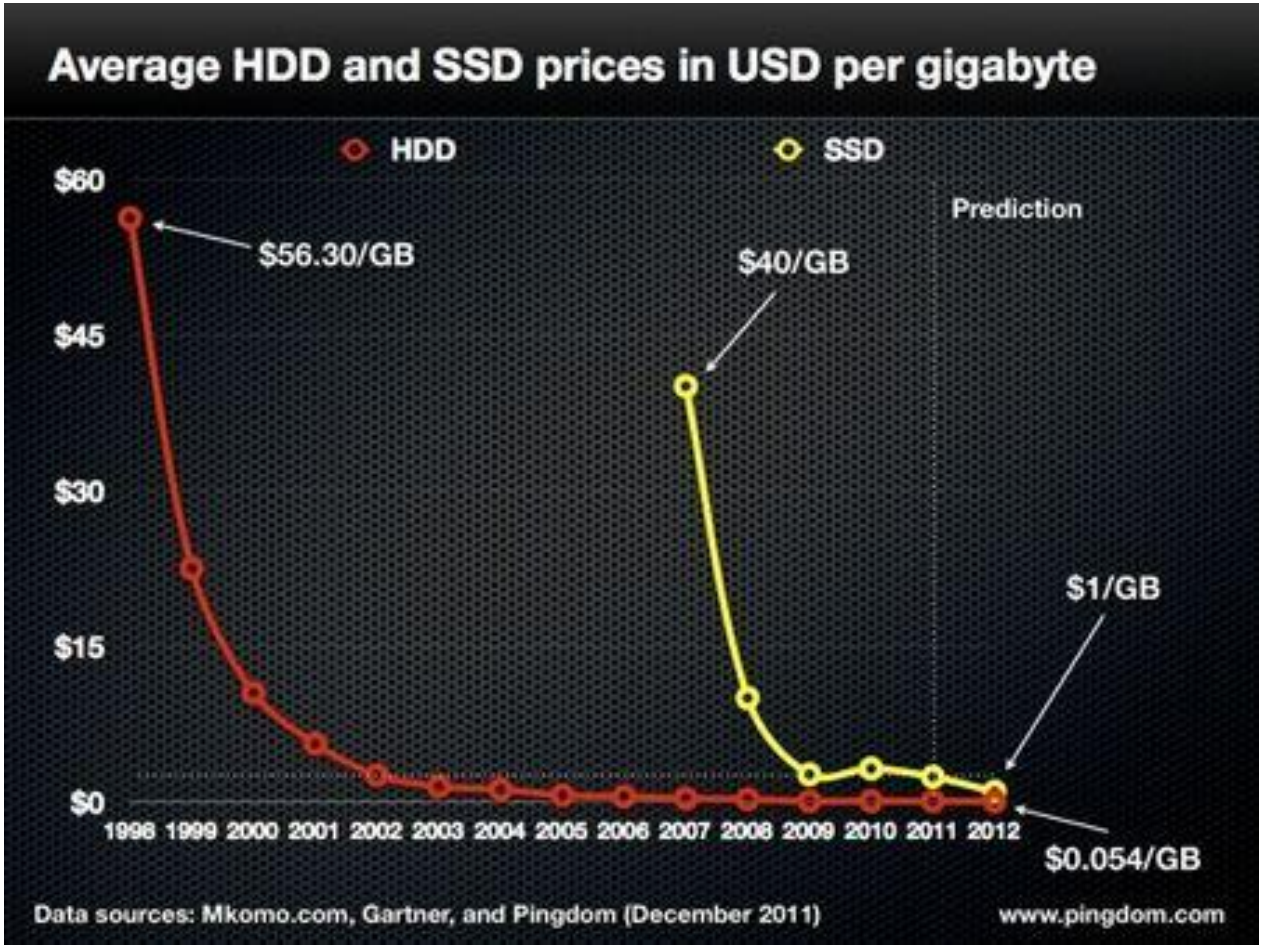
- Si lavora sul miglioramento di flash endurance:
  - Anobit (Apple) promette fattore fino a 40;

- Phase Change Memory (PCM) SSD è stato presentato da UCSD nel 2011.
  - oltre a prestazioni più elevate, endurance verso 100M di cicli;
  - errore catturabile nel momento della scrittura;

- Memristor in vista: in ottobre 2011 HP ha annunciato 18 mesi di lead time per produrre il primo chip basato sui memristor in sostituzione di flash per SSD. Wattaggio basso, endurance.

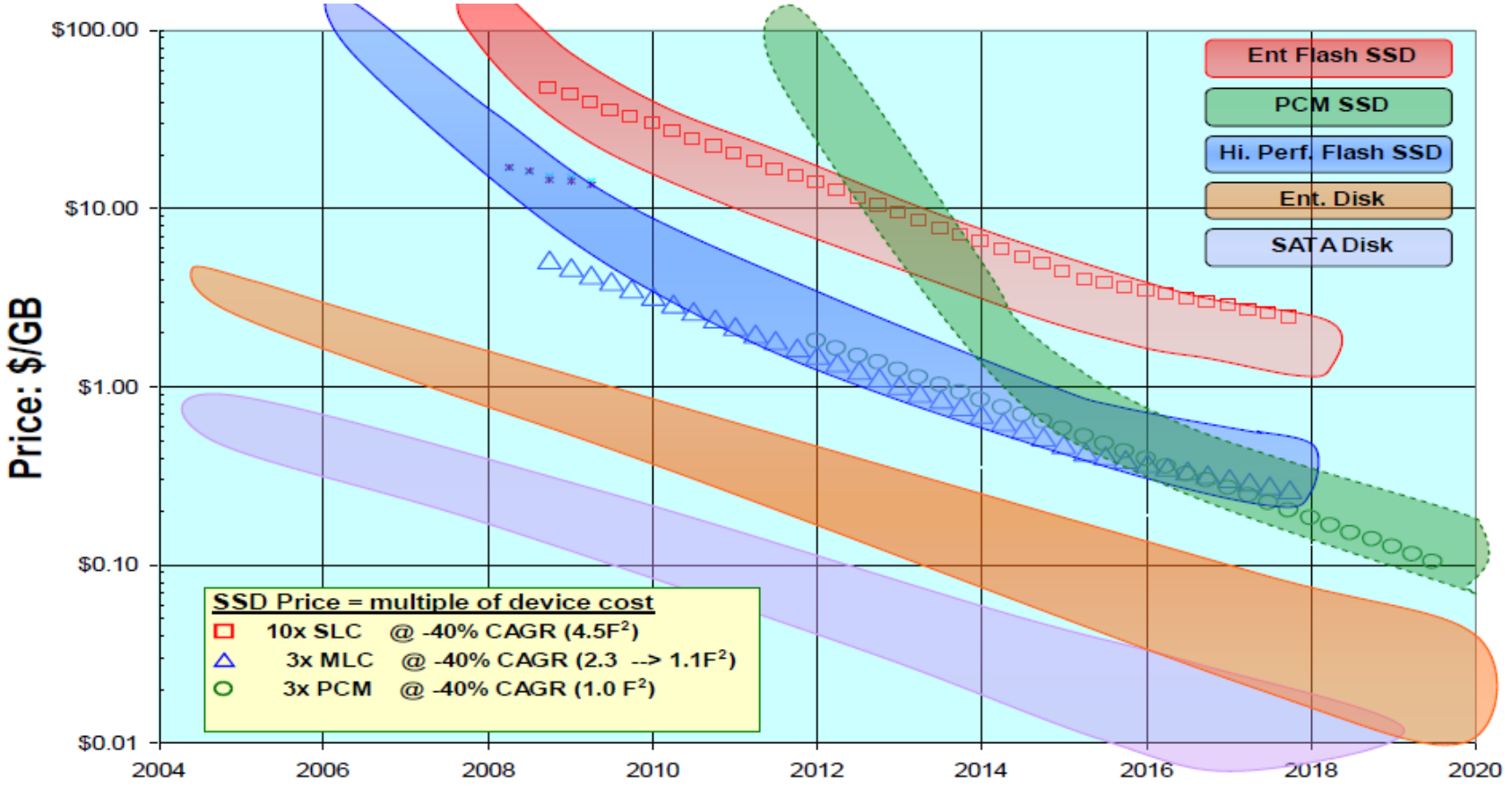
- Dinamica dei costi: meno di 750 E /TB per il consumer MLC (fine 2012)

# Dinamica dei costi SSD MLC




- **Anche se in partenza i costi di SSD MLC sono calati molto rapidamente, i prezzi sono ancora molto lontani da quelli di HDD.**

# Dinamica dei costi disco fino a 2020 (IBM 2012)




- **PCM SSD è una delle tecnologie attivamente promossi da IBM. Secondo loro, entro il 2018 il mercato sarà dominato dai PCM SSD (e Memristor?). Si prevede comunque che il prezzo di 1 GB di HDD sarà almeno 10 volte più basso rispetto agli SSD.**

# Limitazioni degli HDD

- Nell'ultimo decennio il rapporto tra i Flops/sec erogabili da una CPU e gli IOPS ottenibili con un HDD è cresciuto a dismisura. Più precisamente, dal 2001 fino ad oggi gli IOPS sono raddoppiati, mentre gli Flops/sec erogabili sono cresciuti fino a 16 volte.

- Le CPU stanno diventando I/O-starved. Il problema è però affrontabile grazie ai sistemi aggreganti RAID e all'utilizzo delle cache SSD veloci di diverso tipo.

# Utilizzo degli SSD


- Gli SSD si piazzano logicamente nella cosiddetta «flash zone» (piccoli volumi, alte prestazioni).
- Per ogni utilizzo concreto, oltre le prestazioni va sempre considerato l'endurance effettivo calcolando il TBW;
- Due esempi recenti (CERN):
  - Nodi batch con un paio di dischi SSD MLC da 300GB (prefetch e analyse)
  - Facebook/Flashcache su SSD SLC sui server AFS

# Cloud disk 2012

- Più di cento provider, molti di loro non possiedono lo storage fisico ma lo acquistano dai provider grandi (tipo Amazon).

- Tra le realtà più grosse: Dropbox.
  - 50 milioni di utenti (96% usano meno di 2GB a testa)
  - 220+ PB online su Amazon Simple Storage service (S3)
  - Costi: fino a 2GB gratis, poi 1500 Euro/TB/anno

- Amazon Cloud Drive: 5GB gratis, poi 770 Euro/TB/anno;
- LiveDrive: 85 Euro/TB/Anno;

- Utilizzo: disaster recovery D2D2C per i piccoli volumi di dati critici. Usare più di un provider per fare le copie indipendenti. Valutare la rete, resilience dello storage sottostante, solidità della ditta etc. Criptare tutto prima di salvare.

- Amazon S3 nativo: circa 500 E/TB/anno (>5 PB), è un business grande. Oggi nascono gli apparati all-in-one per i provider. Un esempio: DDN WOS (Web Object Scaler); supporta S3, SNIA CDMI, scala fino a 28PB.

# NFS, AFS, GPFS, Lustre, HDFS/GlusterFS..

# NFS 2012

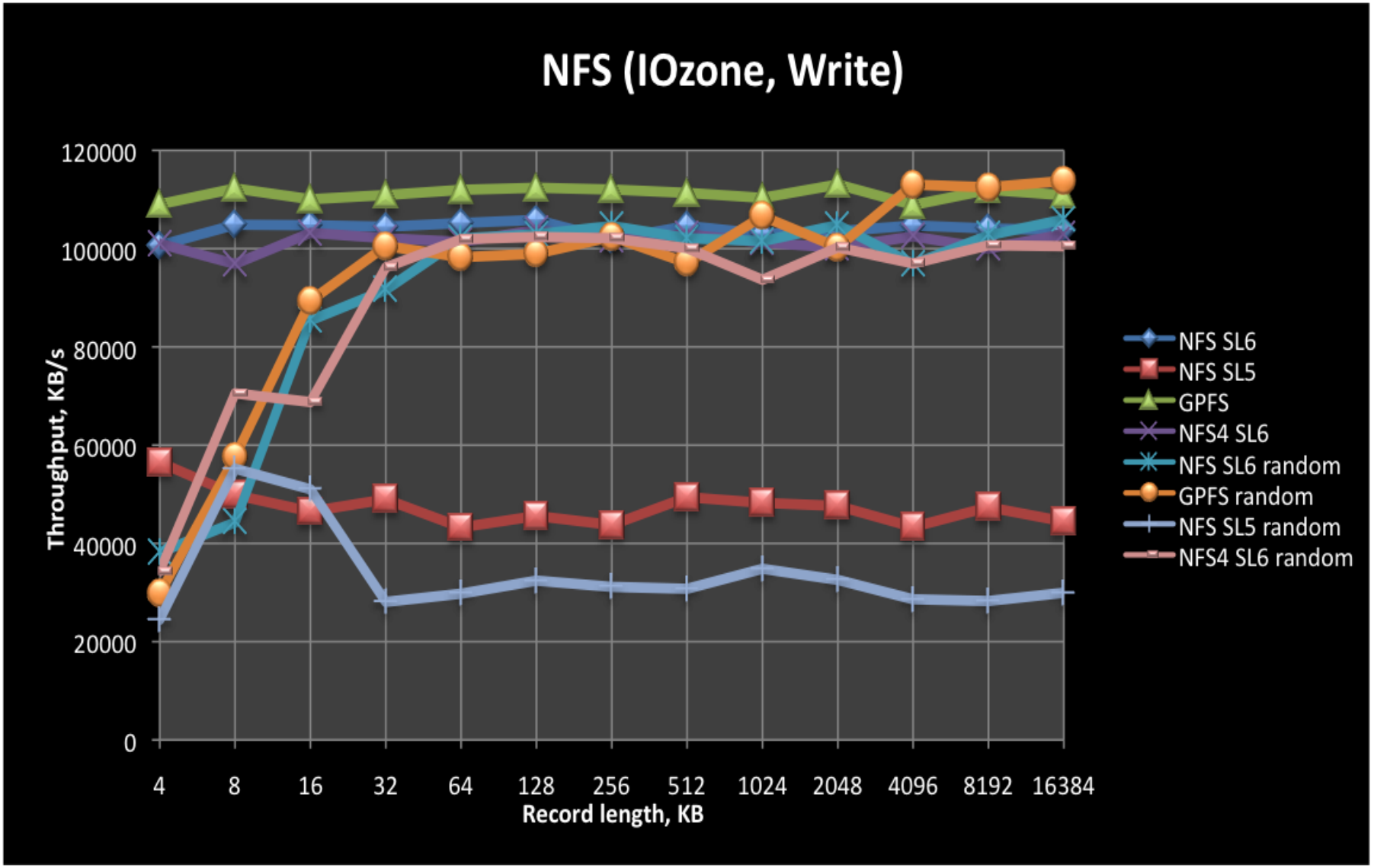
- NFS v4 è nel kernel a partire da 2.6
  - Server e client integrati nel RHEL 6.2;

- Prime prove nel OpenLab di CNAF (vedi la slide successiva)
  - GPFS esportato con NFS v3(SL5) e v4(SL6);
  - Netto miglioramento delle prestazioni di scrittura su SL6;

- NFS v4.1/pNFS (da kernel >2.6.31)
  - RHEL6.2, 6.3 client: supporto solo di protocollo «file»;
  - Fedora 16 client: tutti e tre: «file», «block», «object»;
  - Linux pNFS server è ancora in via di sviluppo;

- I vendor che supportano pNFS (da verificare!):
  Panasas(object), NetApp (file), IBM (GPFS file),
  EMC (block, HighRoad), Oracle (file)

# Prestazioni di scrittura, NFS3 (SL5) vs NFS4 (SL6)

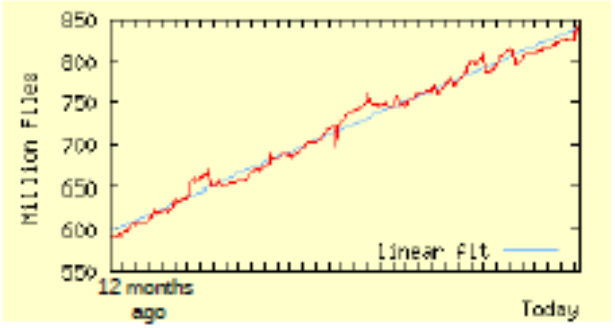


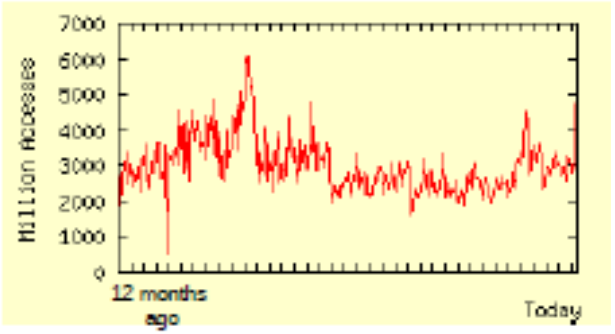
# AFS 2011-2012

- Versione 1.6 dall'autunno 2011
  - Demand attach fileserver, keyring PAGs;
  - Più thread per fileserver, migliore rintracciamento dietro NAT;
  - Più opzioni per la gestione della cache, supporto tmpfs;
  - Eliminazione di varie race conditions, pulizia sorgente;
  - Prestazioni simili a 1.4.x, almeno per alcuni di HEP use case;
  - Baco fs potenzialmente pericoloso in 1.6.0 (risolto in 1.6.1);
- AFSOSD
  - Integrazione con la sorgente mainstream via librerie shared;
  - Osdserver farà parte di vlserver;
  - PSI prossimo deployment site;
- CERN AFS upgrade (seguono un paio di slide copiate da A.Wiebalck)
  - Quota AFS da 100GB per l'utente;
  - AFS «Storage Unit»: fault tolerant, economico, performante

# AFS @ CERN

- Service provides networked storage to CERN users
  - >30'000 home directories & ~300 project spaces
  - high availability, daily backup, security, access control, quotas, monitoring, operability, ...
- Service key data
  - ~55 file servers
  - ~850 million files (+250 m/yr)
  - ~55TB of data, ~100TB quota
  - ~65'000 volumes
- Service activity
  - 10'000 CERN clients
  - 5'000 off-site clients
  - 5'000 active users/week
  - ~3 billion accesses/day
  - ~300 million reads and writes/day



# CERN AFS Storage Unit

- Hardware Setup
  - 2 servers and 2 trays form a “unit”
  - all disks visible on both servers
  - 16x 2TB NLSAS, 4x 256 GB SSDs
- Reliability
  - JBODs (no h/w raid controllers)
  - s/w RAID across arrays
  - “volume take-over”
- Performance
  - make use of SSDs to compensate larger disks: FACEBOOK's flashcache


Storage Unit (SU)

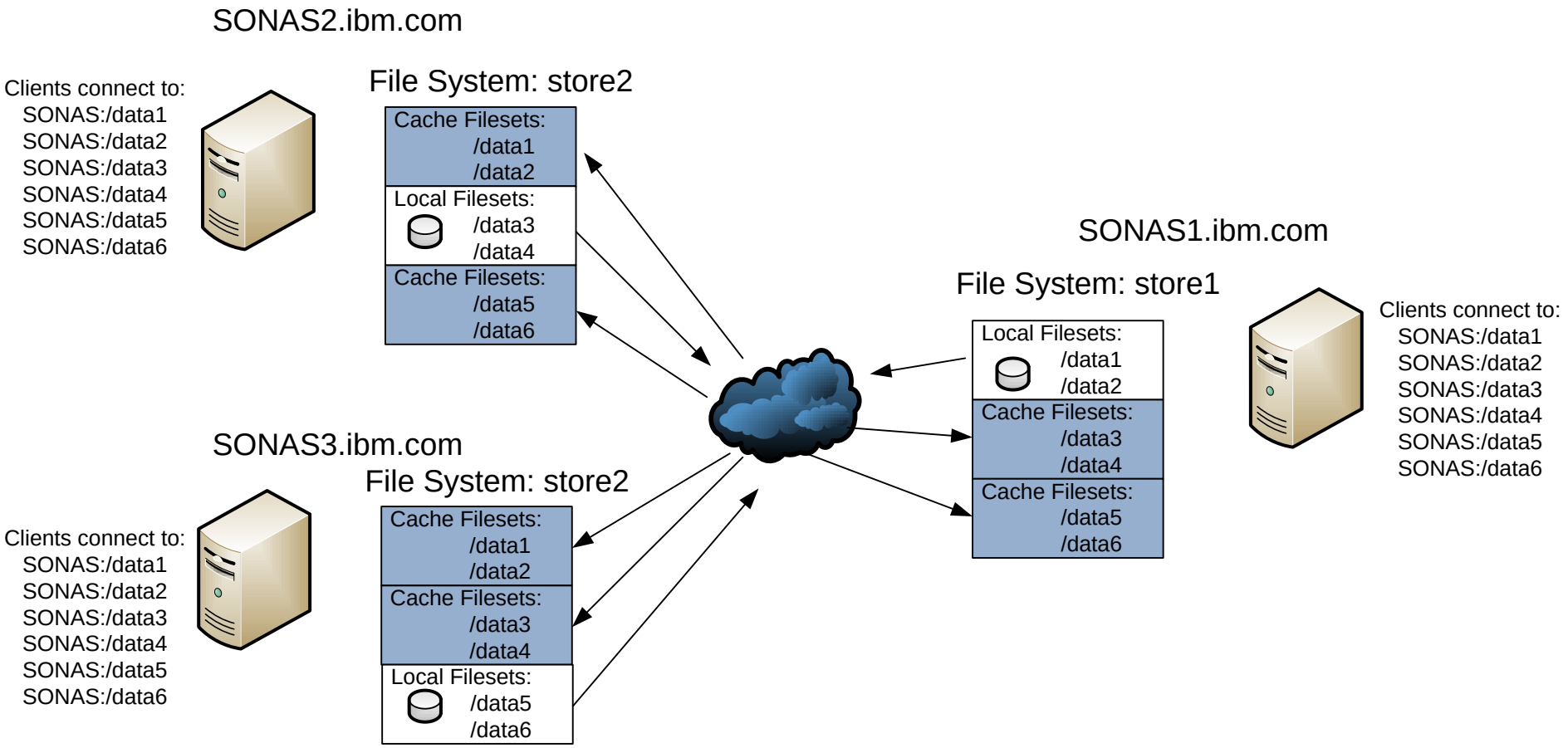
# GPFS 2012

- Versione 3.5 a partire da aprile 2012
  - Active File management (possibilità di global name space con caching, vedi la slide successiva);
  - Supporto IPv6;
  - Windows port nativo;

- Licensing:
  - Costi sulla base del numero e della tipologia dei core;
  - Licenza sito scontata è negoziabile (INFN è già coperto);
  - Con gli NSD directly attached sui client bastano pochi server;

  Un tentativo di stimare il prezzo per 4 server e 100 client (tutti 2xE7)
  8 «Lg 1 CPU Server» + 200 «Lg 1 CPU Client» → 14600 E (ok?)
  Rif: Google per ENUS209-105.PDF

# Es: GPFS global namespace con caching



# Lustre 2012-2013

- Attuali versioni: 1.8.7 e 2.2.0
  - Whamcloud: sviluppo, manutenzione sorgente, supporto;
  - Comunità molto attiva, bug tracking, forum etc;
  - Release più frequenti, versione 2.3 è pianificata per settembre;
  - Problemi iniziali 2.2.0: baco «blocker» di Inet (risolto);
  - Ottima stabilità dopo 1.8.7;

- Whamcloud ha pubblicato la roadmap 2012 (vedi la prossima slide).

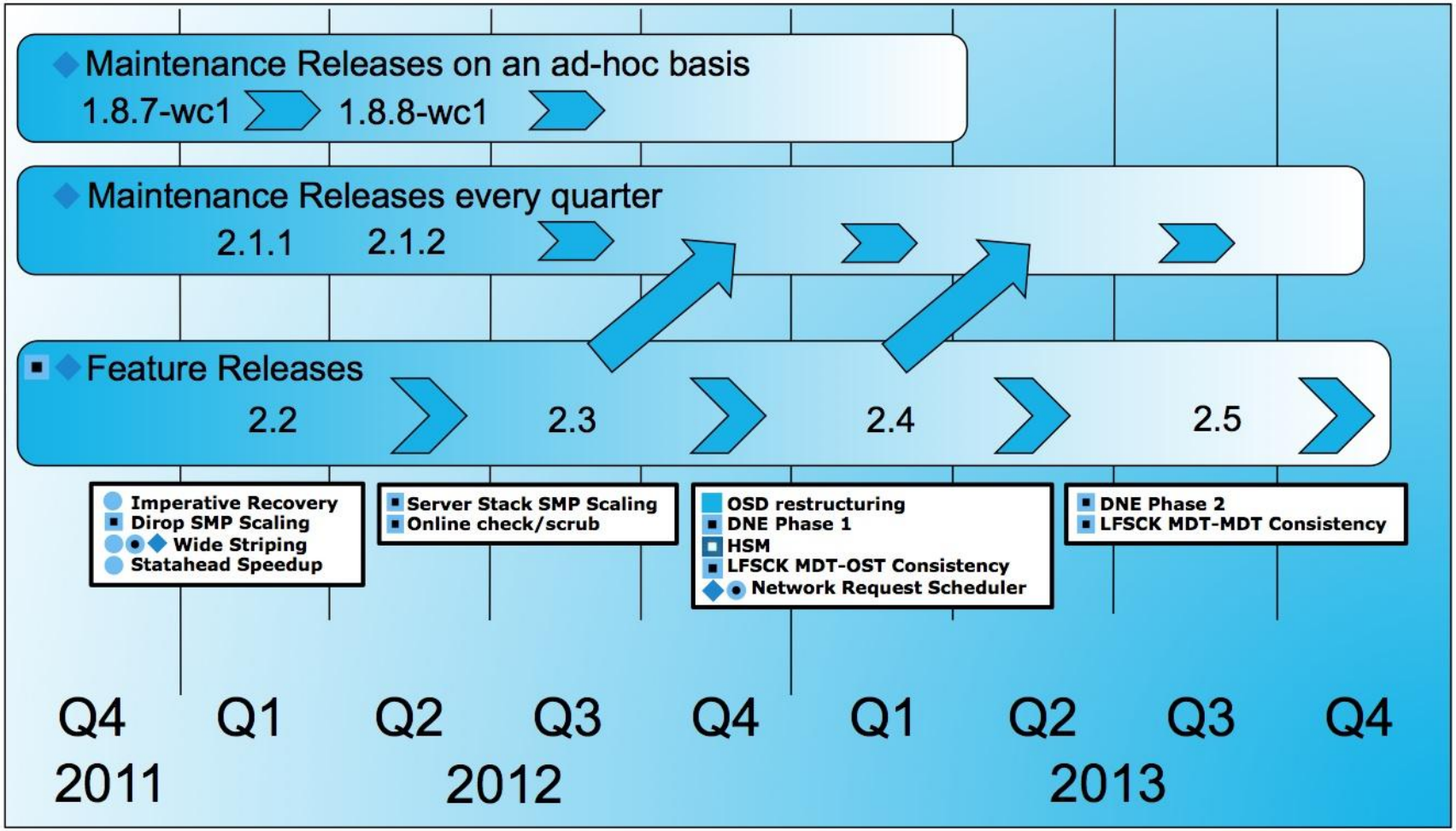
  Le features di particolare interesse includono:

  - OSD restructuring (finalmente ZFS su OSTs e MDTs);
  - LFSCK online check/scrub (distributed repair);
  - Distributed namespace (load balance si più MDT);
  - HSM;

- TWG requirements di OpenSFS edizione aprile 2012 :
  lunga lista di richieste (alcune già nella roadmap)

- Chroma (management suite), costi: da capire

# Community Lustre Roadmap




Sponsor for Whamcloud Development and Releases: ORNL OpenSFS LLNL Whamcloud

Third Party Development: CEA Xyratex

# HDFS e GlusterFS 2012

**HDFS**

- **Attuale versione stabile: 1.0.2**
- **Versione 2.0 (beta):**
  **- Separazione netta di name space e block storage;**
  **- Snapshots;**
  **- Supporto per name node HA;**
  **- MapReduce NG;**

**GlusterFS (Gluster ora appartiene a Red Hat)**

- **Attuale versione stabile: 3.2**
- **Versione 3.3 (beta):**
  **- Accesso unificato file/object;**
  **- Hadoop hooks;**
  **- Nuovo tipo di volume («repstr»): replicated+striped(+distributed);**
  **- Rebalance può migrare i file aperti;**
  **- Remove-brick può migrare i dati sui brick rimanenti;**
  **- Granular locking, proactive self-heal, quorum enforcement**

# Costi 2012  Prêt-à-porter

# Alte prestazioni:
# un HPC cluster di 1000 nodi con Infiniband

- DDN SFA12K-40 (1 rack, IB/FC, 14KW, 40GB/sec, 2PB netti) - 650 KE

- Con GPFS e gli NSD directly attached sui client via IB/SRP possono bastare anche soli 4 server per cluster management e quorum;
  4 x E9128-2 (2xE5-2650, 32GB, IB 4x) – 23KE (*)

- Licenza GPFS per 4 server e 1000 client (stima imprecisa):
  4x2700 E + 1000x36 E = 46KE

- Totale: circa 720 KE per 2PB netti, **360 KE / PB**

- Per ottenere le stesse prestazioni con Lustre il costo sarebbe molto più alto in quanto servirebbero tanti server (un server Lustre è in grado di erogare non più di 1GB/sec su Infiniband).

(*) Origine prezzi server – E4.

# Prestazioni medio-alte:
# un cluster di 1000 nodi con il turnover di 1TB/nodo/giorno

- Il throughput richiesto è di circa 12 GB/sec ed è ottenibile con 15 server Lustre; 15 x E9128-3 (2xE5-2650, 32GB, 10G) – 78 KE

- 2 PB netti dopo erasure coding con il fattore 1.2 richiederebbero 600 dischi da 4TB (E 7200 RPM). Il costo sarebbe di circa 330 KE.

- I 600 dischi possono essere messi in un 10 enclosure SAS; 10 x (DataON DNS-1660, 60 slot, 4RU) = 85KE

- Totale: 493 KE per 2PB netti, **247 KE / PB**

# Prestazioni nearline (capacità)

- 2 PB netti dopo erasure coding con il fattore 1.2 richiederebbero 800 dischi da 3TB (C 5400RPM). Il costo sarebbe di circa 120 KE.

- Gli 800 dischi possono essere messi in un 14 enclosure SAS; 14 x (DataON DNS-1660, 60 slot, 4RU) = 120KE

- Alla fine servirebbero anche soli 7 server a basso costo con GlusterFS e una scheda SAS: circa 18KE

- Totale: 258KE per 2PB netti, **129 KE / PB**

# Discussion