designed. engineered. results.

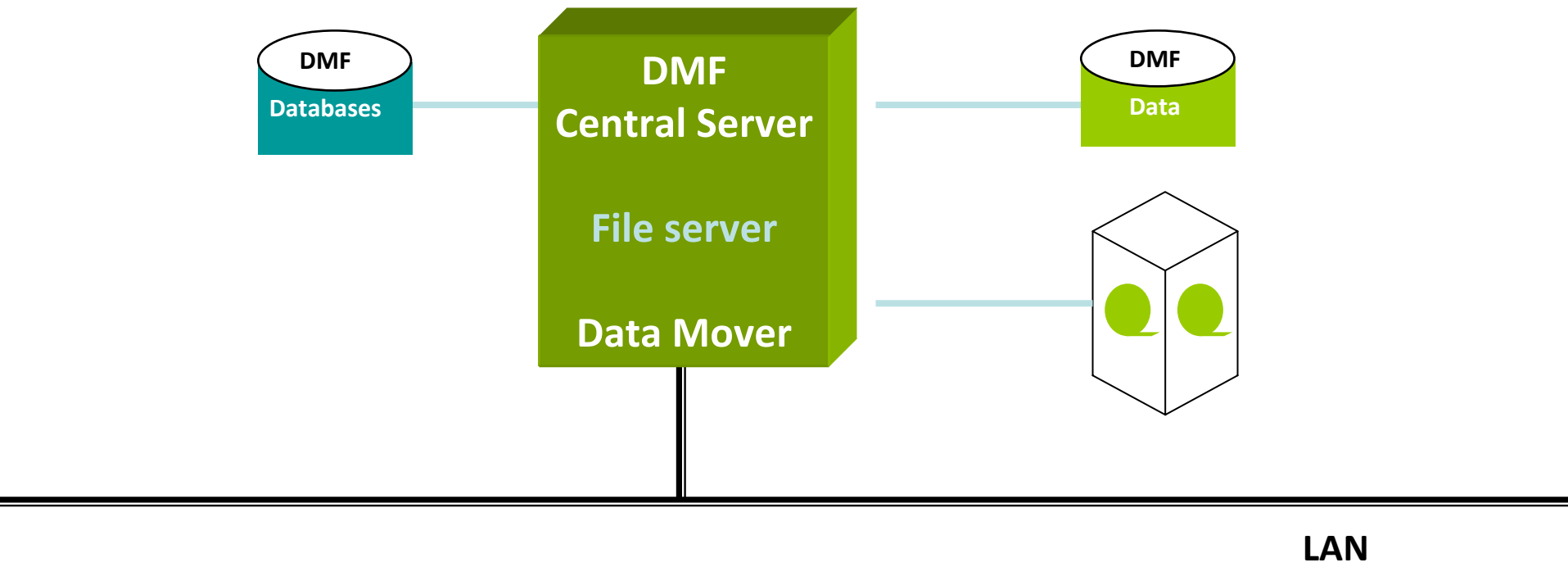
Parallel DMF

Agenda

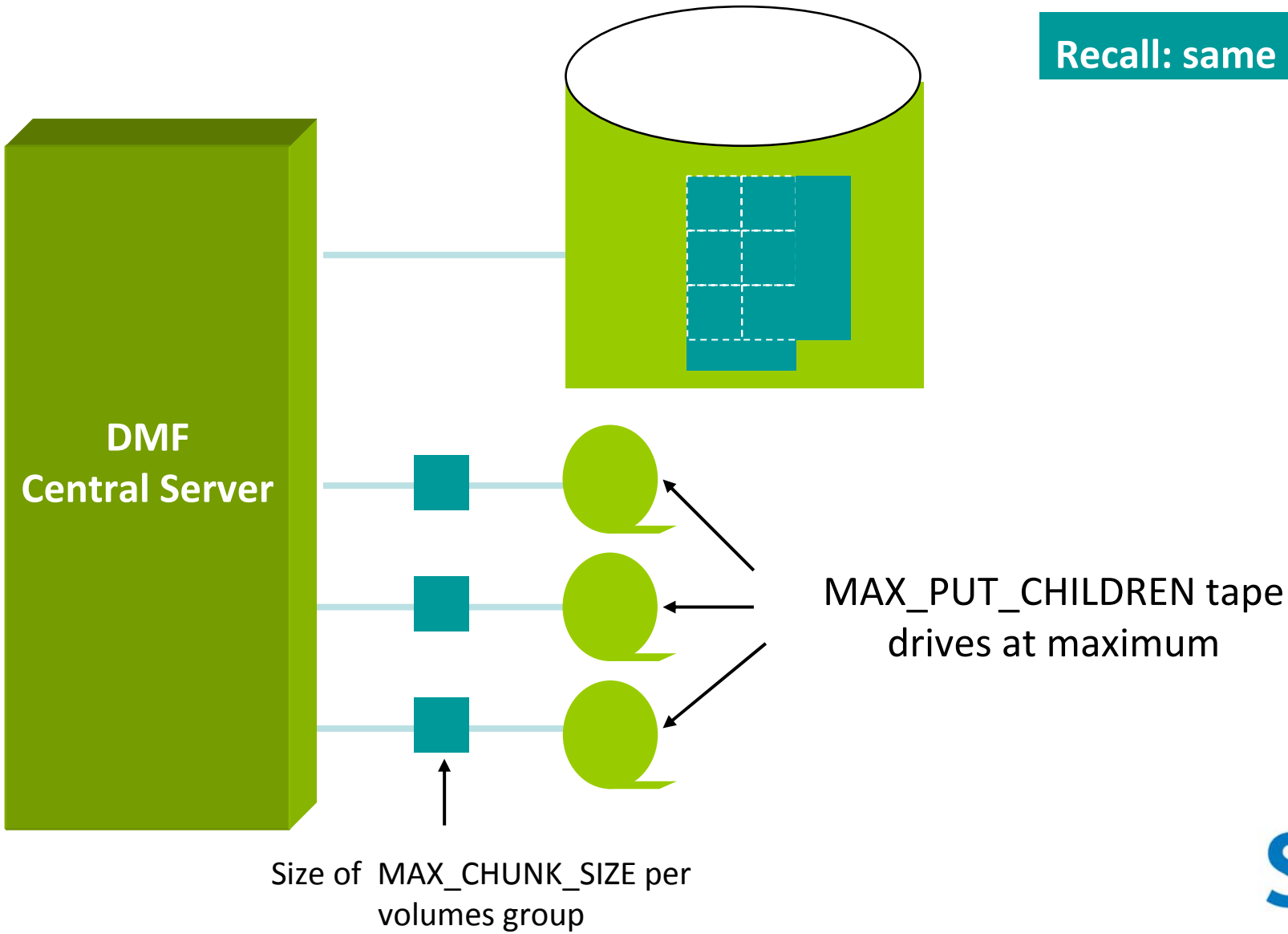
- Monolithic DMF
- Parallel DMF
- Parallel configuration considerations

Monolithic DMF

Monolithic DMF



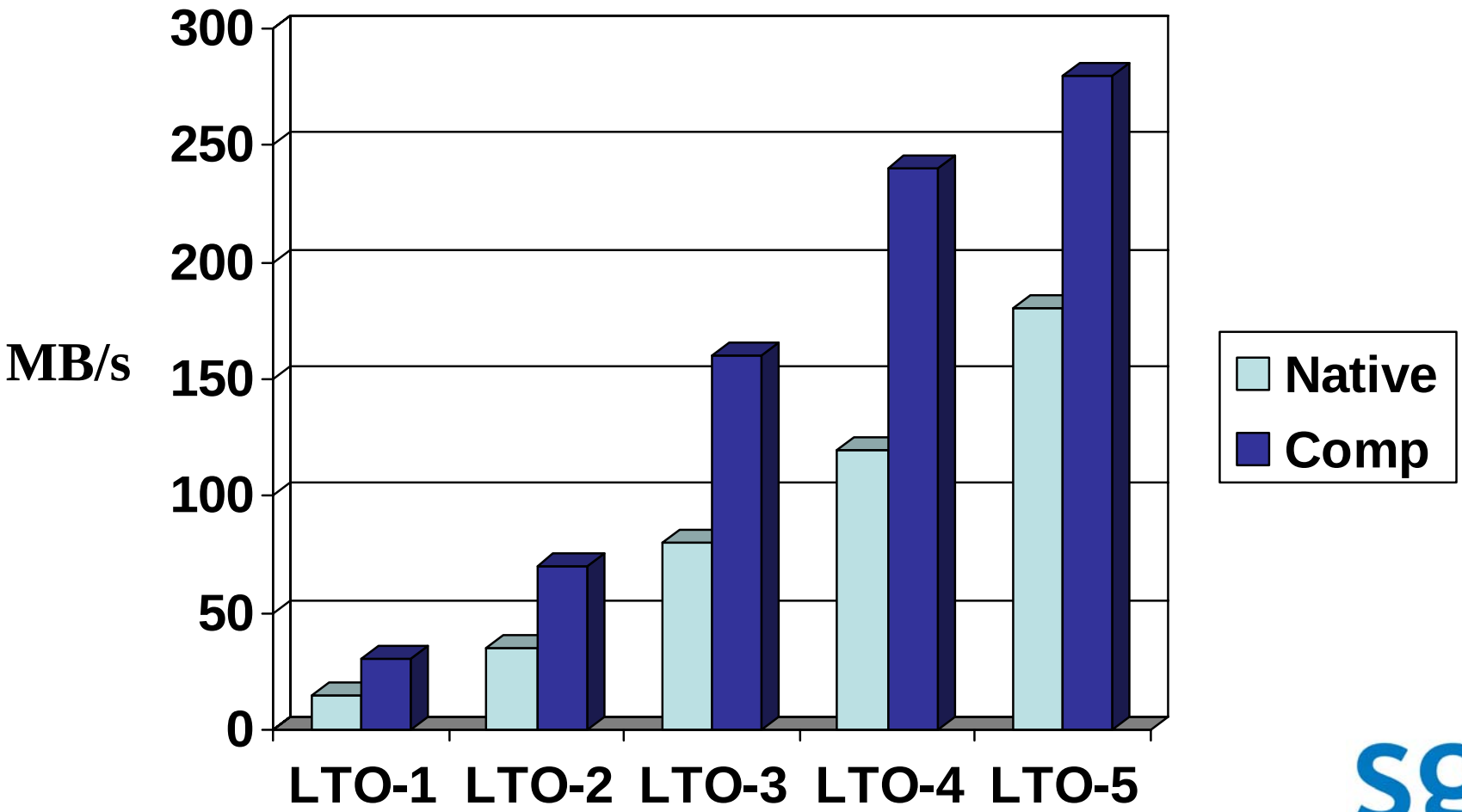
Data Movement Parallelism



Parallel DMF (pDMF)

Why pDMF – technology upgrade

Increasing per drive transfer rates



Why pDMF - single server limitation

- Single server will bottleneck at some point
  - CPU cores do not get equal bandwidth to all memory
  - buffer placement creates bottlenecks
  - DMF, FS, page cache, tape I/O, …
    - Not all software layers optimize buffer placement
- No longer possible to move all the data with a single system

sgi

pDMF Benefits

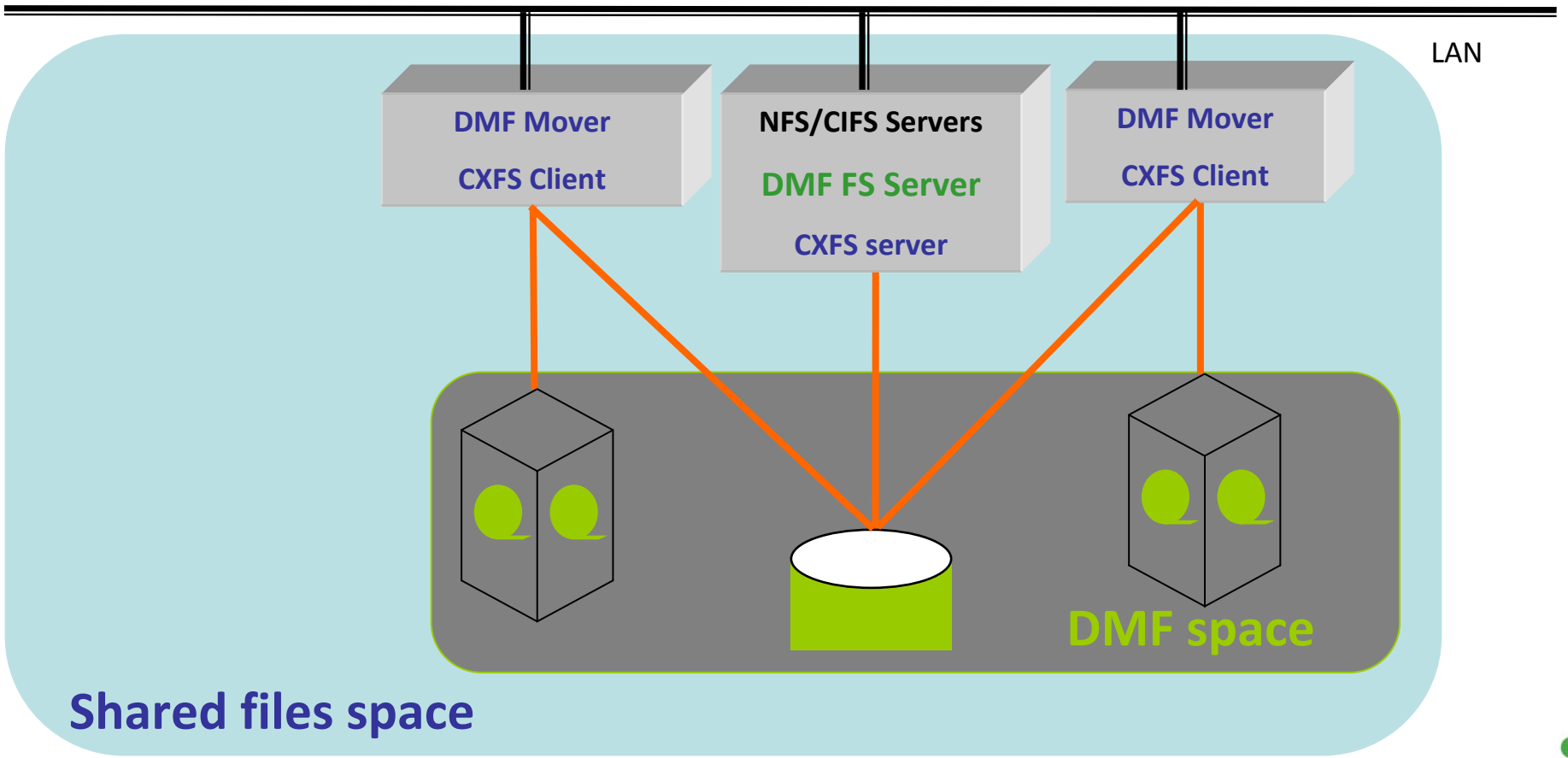
- Scale throughput to/from the tape drives
  - Add mover nodes as needed
- Provide redundancy
  - Loss of a data mover node
  - Loss of a tape path on a data mover node
- Lower cost
  - x86_64 hardware
  - small central server
  - HA solution

DMF 4.0 is Parallel DMF

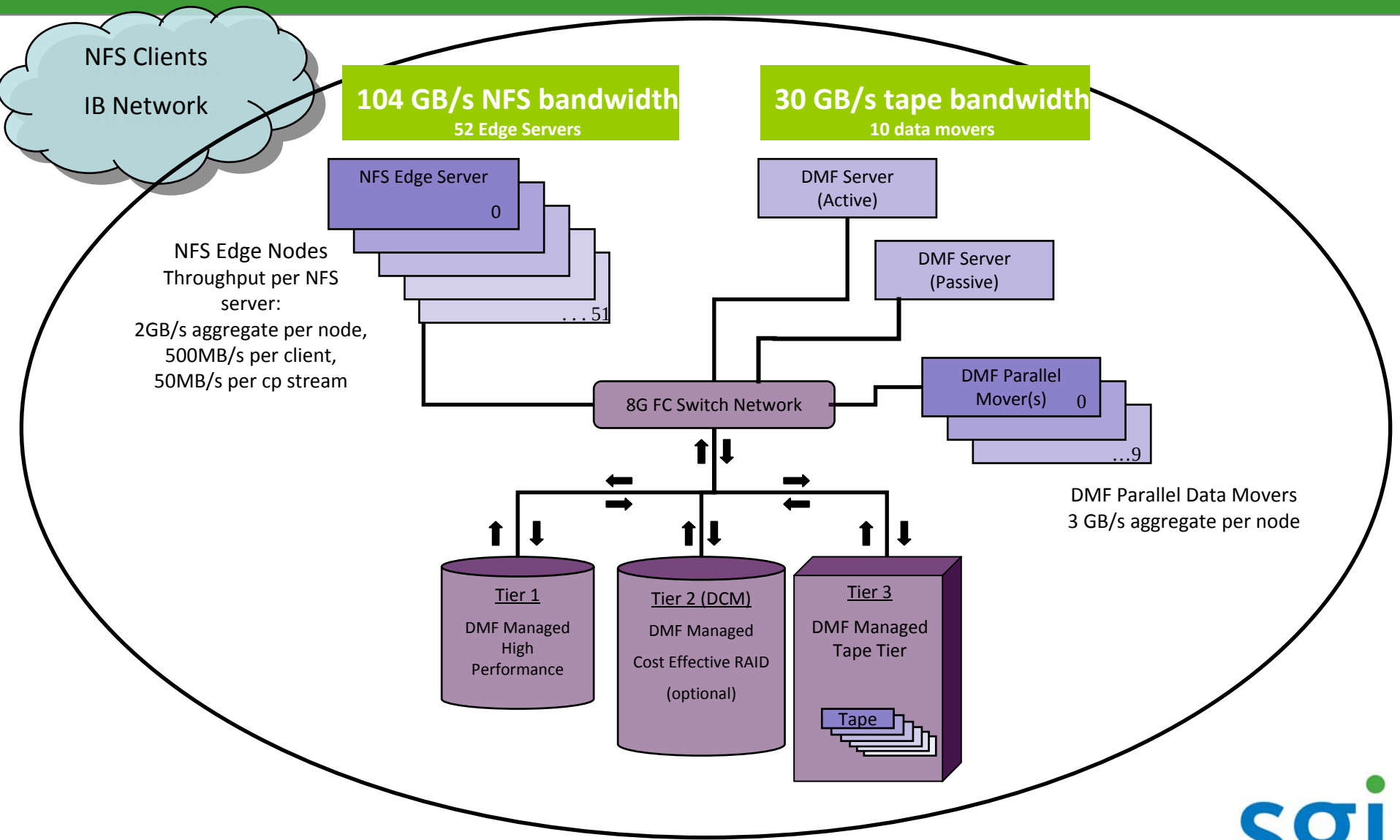
- Current release DMF 4.4.1/5.0
  - 4.4.1 (sles10sp3)
  - 5.0 (sles11)
- Runs in monolithic configuration
  - traditional way in which DMF has run
  - DMF's data movers run on the DMF server
- Runs in parallel configuration
  - Scalable way in which DMF 4.x can run
  - DMF's data movers run on dedicated nodes
- Same DMF code base for parallel & monolithic

sgi

Parallel DMF



SGI Scalable Storage Architecture



Data Mover Platform

- Select hardware targeted for data movement
- x86_64 better than ia64
  - Both price and performance

- Best price/performance mover node
  - C1103 (Rackable server line)
  - 1U / 2 socket / 96 GB DDR3
  - 2 PCI-E x16 or 4 PCI-E x8

Summary of architectural advantages

- I/O Scalability
- HA Redundancies
  - Data Movers
  - DMF Server
  - NFS Edge Servers
- Flexibility - choose platforms according to their task
  - DMF Server
  - Data movers
  - NFS Servers
- Utilization of low cost x86 servers

pDMF Considerations

Configuration Considerations

- DMF server needs access to some tape drives
  - backup, dmatsnf, dmatread.. run only on the server
  - Server can optionally be a data mover node too
- Not all data movers need to see all tape drives
  - Best practice: 2+ movers can see each tape drive
  - If the last mover goes down, tape drive won't get used
  - If no mover can see the tape drive the cartridge might require admin intervention to become usable again (just as today with monolithic DMF)

Configuration Considerations

- DMF user and admin filesystems need to be CXFS filesystems
  - server and mover nodes need fast access to the filesystem
  - Movers need to be able to dmapi reads/writes to the filesystems
- Disk Cache and Disk MSP
  - All I/O currently done on DMF server

Library Robot scheduling

- Openvault mounting service required for parallel configurations
- Released version has a limitation in multi-armed ACSLS controlled libraries
- June release eliminates this limitation
  - asynchronously schedules as many cartridge movements as ACSLS has tape drives

Tape drive scheduling

- General Rules
  - Choose a drive in the same library bay as the cartridge if possible
  - Balance load across mover nodes
  - Balance load across HBA ports within mover nodes
  - If several choices, choose least used drive for wear-leveling

- per mover config param for FC port speed
  - all HBA's should have the same Gb/s port speed
  - chooses a port with the most remaining bandwidth
- per mover config param for total bandwidth
  - each mover can have different throughput capabilities
  - chooses a mover with the most remaining bandwidth

Tape Merge Considerations

- Socket merges
  - Merges may be done between mover nodes
  - Network bandwidth generally not enough to keep up with modern tape drives
- When do socket merges occur?
  - files too large to fit in the cache dir filesystem
  - files larger than a configurable size
    - default is files larger than 25% of the the cache dir
- Recommend tape merges using cache dir

SAN Configuration

- Use FC zoning to isolate tape drives
  - Prevent unexpected tape rewind with risk of data corruption
  - Persistent reserve in OpenVault to come
- Disk and tape should be on different HBA ports
  - CXFS may fence the disk HBA port

www.sgi.com