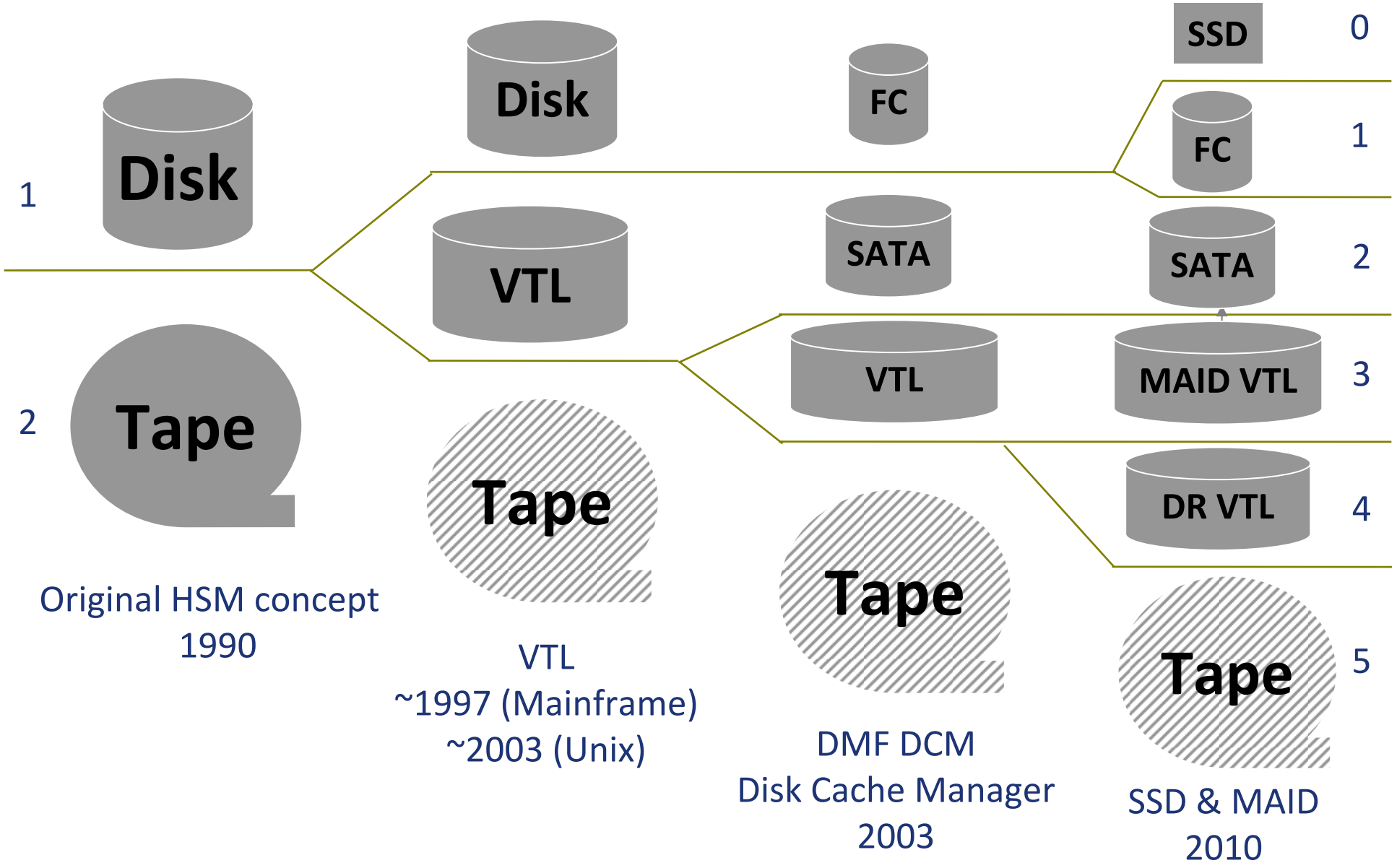
Using VTL and MAID with DMF

David Honey
SGI Consulting

It’ll end in tiers!



What is VTL?

- A Virtual Tape Library is disk presented as tape by a layer of software
- Tape applications need no modification (Non-disruptive way to add disk)
- Because it's an array of disks, it supports random IO and is fast
- Can include de-dupe (post processing)
- Enhances DR by replicating over FC or IP networks and improving recovery speed
- Enhances device management
- Removes tape streaming issues

But disk is more expensive than tape…

VTL is used to get more performance from applications that use tape

So VTL is more expensive than tape…

Yes, performance comes at a price

What is Copan MAID?

- A Massive Array of Idle Disks is RAID protected SATA disk storage that only spins when it is accessed
- MAID uses extremely dense packaging of high capacity drives
- Power Managed RAID ensures no more than 50% of disks are spinning
- In standby mode 2.6% of drives are spinning (0.8 – 1 shelf to 2.7kW – 8 shelves)
- Fully operational, a rack uses between 2.7 and 7.5kW (3 to 50% spinning)
- MAID is most commonly presented as VTL. Native option (DMF integration with native MAID is under consideration)
- Faster than tape
- At least half the power of RAID

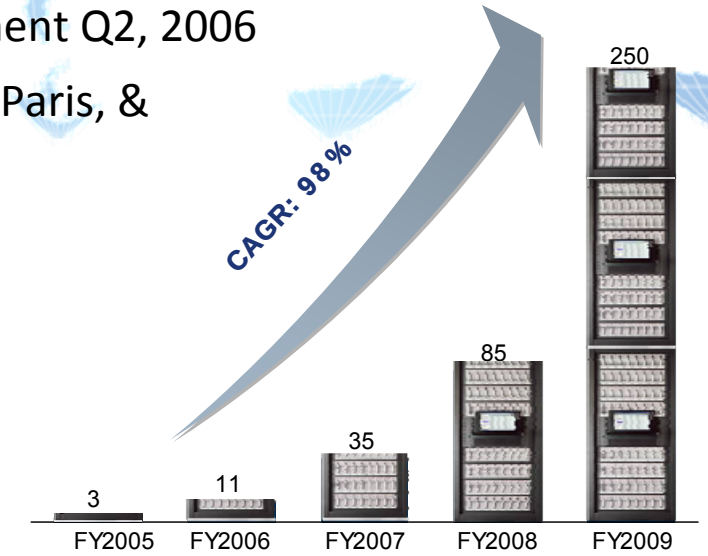
But you told me tape was green, why would I use MAID …

MAID is designed for tier 3/4, Tape for tier 5 and RAID for tier 0/1/2

COPAN Corporate Overview

Organizational Snapshot

- Founded 2002
- Corporate Headquarters, in Longmont, CO
- 66 employees, WW
- FCS Q1,2004, First International shipment Q2, 2006
- International Sales offices in: London, Paris, & Cologne
- Distribution Partners in Japan
- World-class partners



Installed Capacity in Petabytes

Different Types of Data Demand Different Storage

Transactional or Dynamic Data

- I/O intensive
- Small files
- Modest storage growth
- Steady growth rates

E-mail 8KB | Document 80KB | Database 8MB

Persistent Data

Data Protection and Archive Data

- Large files
- Very large storage
- Throughput
- Sequential
- Explosive growth

Backup 10MB | Replication 20MB | Maps 60MB | Video 300MB | Imaging 48GB

Vaulted Data

- Offsite
- Copy of copy
- Sequential
- Compliance

Transactional data → **Persistent data** → **Vaulted data**

Relative proportions of data in the typical enterprise

Within 30 days the majority of transactional data becomes persistent data

sgi®

COPAN Storage

Copan MAID Engineered for Tier3/4

- Disk-based storage for
  - Backup…based on the restoring of data and cost
  - Archive…based on the scalability and cost
- Access to all digital assets
  - The increased value on readily available assets
    - Automated retrieval by the same applications that archived the container
      - e.g. Providing a scalable, low cost file storage product that utilizes the same path on which the data was created. The same file would be retrieved from the same volume/directory/fail path....and retrieved year ago.
  - The burden (cost) of storing, managing and retrieving
    - Reduce the man-hours needed to catalog and store the data
    - Reduce the operating costs for managing un-accessed archived objects (file, volume, tape)
- Provide a configuration to match the cost with
  - Access profile
    - %archive vs. %retrieval vs. %updates
  - Bandwidth profile
    - High “watermark” for archive and/or retrieval
  - Processing profile
    - Cataloging, Indexing, Searching and Migration services
  - Data Management profile
    - Migration, Replication, and Disposition rules

SGI MAID Hardware

- 14 SATA drives per canister
- Patented canister quick release, quick servicing and mounting scheme to minimise rotational vibration
- Total drive power down extends drive life
- 4.5U high
- Extremely dense!

Disk Aerobics

- **Periodically exercises idle drives**
  - Spins drives at least once every 30 days and performs self test
  - Performed as background task with no impact on I/O operations
  - Updates SMART database
- **Actively monitors and manages drive health**
  - Every 5 minutes monitors a subset of SMART* parameters & stores in database
  - Monitors environmental data
    - Excessive timeouts on single I/O, fails on SMART return status command, spin-up retries, reallocated sectors, start/stops, power-on hours, raw read error rate and seek error rate.
  - Monitors environmental data
    - Disk drive temperature, canister temperature, fan activity and voltage
- **Disk Scrubbing**
  - Background task during routine DISK AEROBICS
  - Identifies potential bad sectors on disk & copies data to new sector on drive
  - Increasingly important as drive sizes increase
- **Proactive failing of "suspect" drives**
  - Suspect drive exceeded predefined thresholds or met specific conditions
  - Data proactively copied to spare disk and inserted into RAID set
  - Suspect drive taken out of service

Disk Aerobics and spin down extend drive life

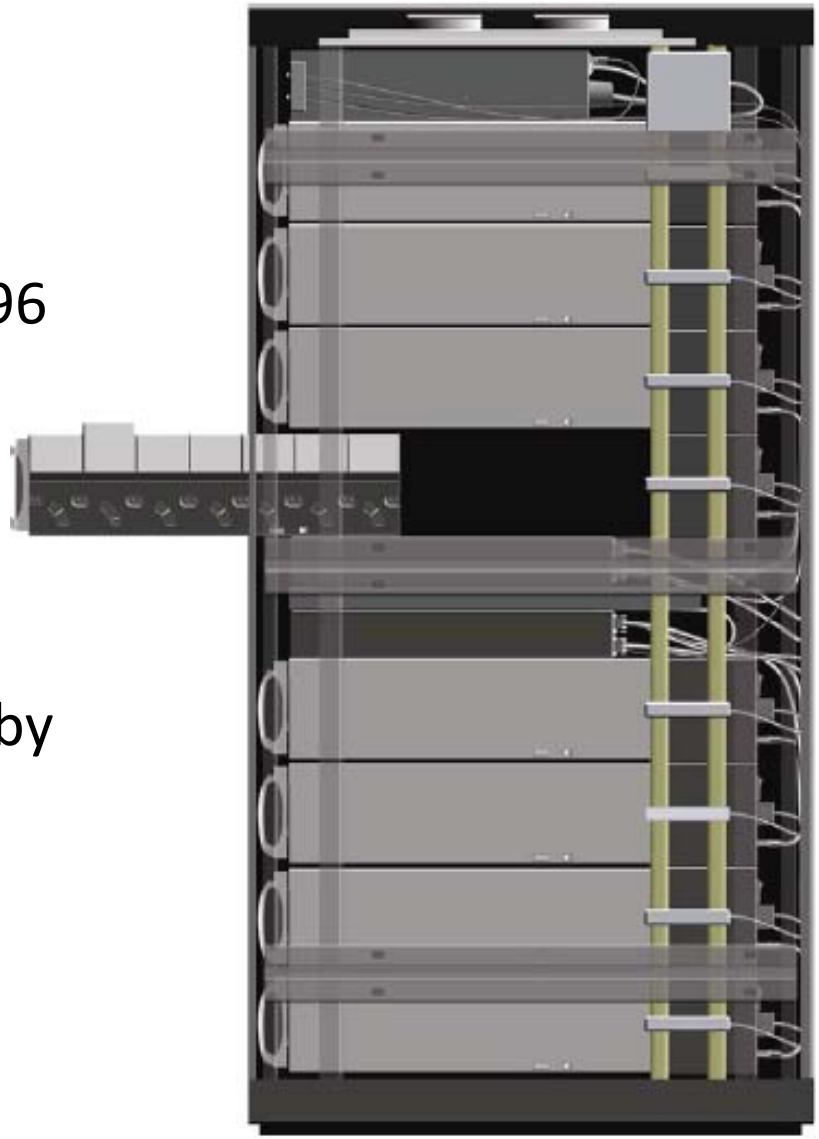
- **Extended drive life and reliability**
  - Compared to standard SATA disks, COPAN has less than ¼ the failure rate
  - Field MTBF: more than 4X SATA disks, more than 2X FC disks
  - Service Life: expect more than 4X
- **Disk Reliability and TCO benefits**
  - Assuming 1000 drives, expect:
    - COPAN: 3 failures/yr
    - SATA: 15 failures/yr
  - Standard SATA platforms have
    - ~5X drive replacements
    - 17 touches versus 1 touch for COPAN
- **Data Reliability Benefits**
  - Fewer failures ⇨ 1/23 Data Loss

MTBF (hrs)	AFR (%)	Disk Specification
8,000,000	0.11%	
5,000,000	0.18%	
3,000,000	0.29%	
2,902,706	0.30%	COPAN - Apr 2006
2,400,000	0.36%	
2,000,000	0.44%	
1,200,000	0.73%	Fibre Channel
1,000,000	0.87%	Fibre Channel
800,000	1.09%	
600,000	1.45%	SATA
400,000	2.17%	SATA
200,000	4.29%	
100,000	8.39%	

600K hrs = 68 yrs
2.64M hrs = 331 yrs

SGI MAID Hardware contd.

- 44U cabinet
- 8 canisters per shelf – (112 drives)
- 1 to 8 shelves per rack means up to 896 drives per rack (upgrade unit is a shelf)
- cf. 600 drives in 45U IS15000 rack with 60 drive enclosures
- 1.8PB per rack (raw) with 2TB drives
- 2.7PB per rack (raw) with 3TB drives (by end 2010)
- Heavy as – 1,448kg (3193lbs) per rack



Power Managed RAID

To manage the storage to a power budget there are three possibilities:

- Configure the application so that it accesses no more than a set number of LUNS
- Artificially limit the inbound workload
- Accept longer service times as IOs are queued

The embedded shelf controller monitors power allocation and determines the next available set of disks (RAID groups).

A maximum of 50% of the drives in a storage shelf are available to the user at any given point in time.

POWER MANAGED RAID software manages the power budget, and determines if there is available power capacity to run DISK AEROBICS in the background.

SGI MAID – Rack choices

8 Shelf Cabinet
400M-000000

6 Shelf Cabinet
406M-000000

Drive / LUN mapping

	Can 0	Can 1	Can 2	Can 3	Can 4	Can 5	Can 6	Can 7
Drive 0	1	1	1	1	2	2	2	2
Drive 1	3	3	3	3	4	4	4	4
Drive 2	5	5	5	5	6	6	6	6
Drive 3	7	7	7	7	8	8	8	8
Drive 4	9	9	9	9	10	10	10	10
Drive 5	11	11	11	11	12	12	12	12
Drive 6	13	13	13	13	14	14	14	14
Drive 7	15	15	15	15	16	16	16	16
Drive 8	17	17	17	17	18	18	18	18
Drive 9	19	19	19	19	20	20	20	20
Drive 10	21	21	21	21	22	22	22	22
Drive 11	23	23	23	23	24	24	24	24
Drive 12	25	25	25	25	26	26	26	26
Drive 13	27 AOR	27 AOR	27 AOR	Spare	Spare	Spare	Spare	Spare

Using 2TB drives;
LUNs 0 thru 26 ~6TB (3+1)
LUN27 2TB (1+1+1)

AOR=Always On Region, RAID1 triple mirror for metadata, eg vol headers

Virtual devices

- 1x Virtual Library per shelf
- Emulation of nearly 50 tape libraries; ADIC, ATL, HP, IBM, Overland, Quantum, StorageTek and SpectraLogic
- 6 virtual tape drives per shelf with 25% spinning
- 12 virtual tape drives per shelf with 50% spinning
- Over 30 tape drives emulated, including; LTO, DLT, Super DLT, AIT, Super AIT, 3590, 3592 and StorageTek 9840, 9940, T10000
- 6TB MAID LUNs divided into 1TB virtual carts
- Virtual tapes match physical tape size in cached VTL mode (best practice)

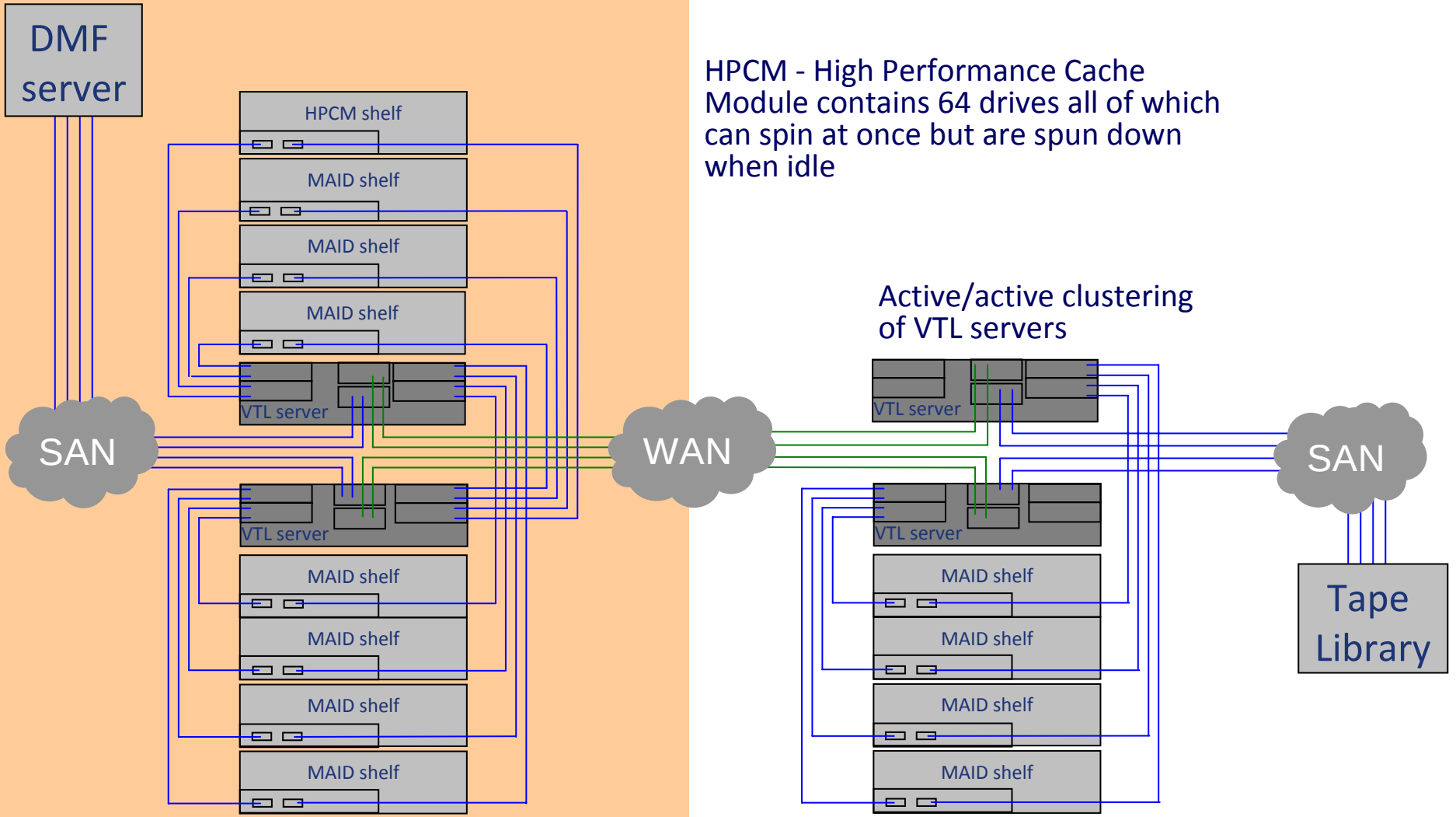
VTL connectivity with DR and HA VTL servers

Primary site

DR site

HPCM - High Performance Cache Module contains 64 drives all of which can spin at once but are spun down when idle

Active/active clustering of VTL servers

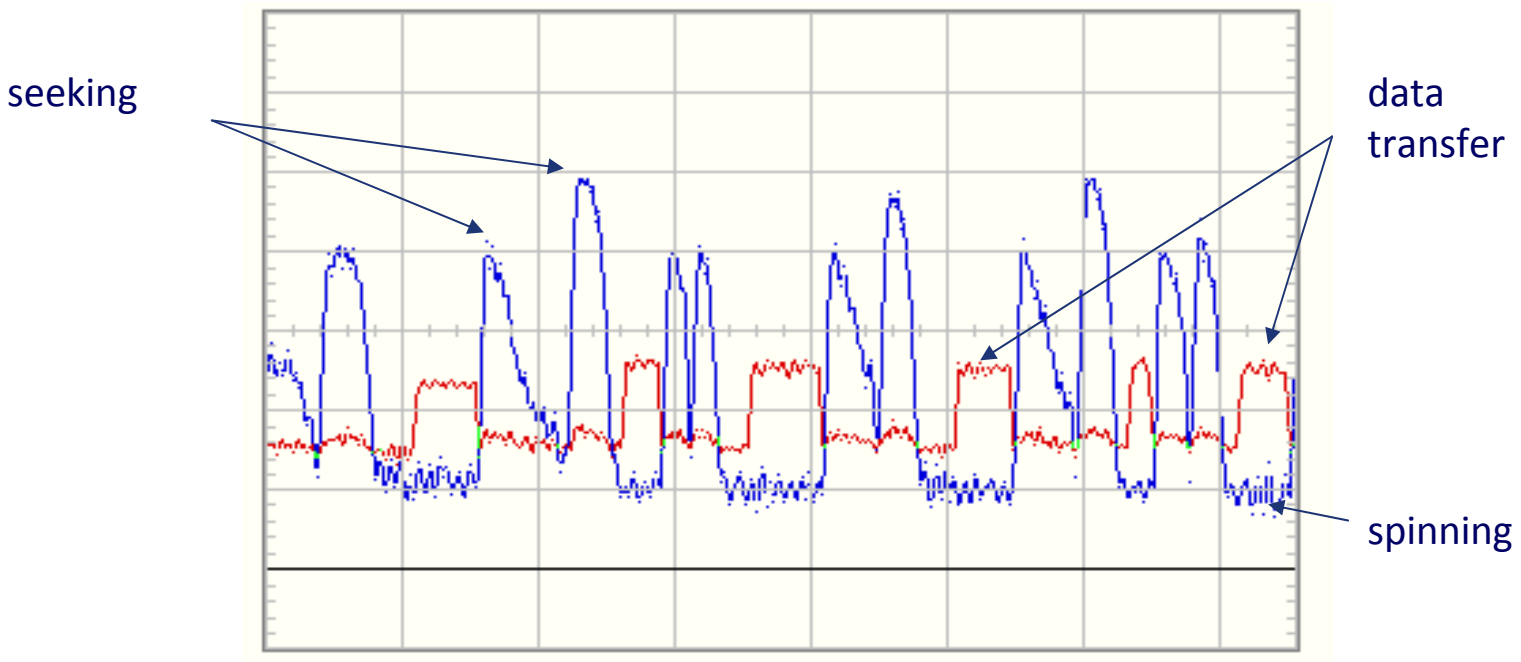


RAID vs MAID vs Tape

	SATA RAID (960TB raw)	MAID (1.8PB raw)	Tape Library (T950) (1.5PB raw with LTO5)
Time to first byte (unmounted volume):	12ms	15s	90s (16s 9840D)
Time to second file (mounted volume):	12 ms	12ms	45s (8s 9840D)
Parallelism:	>1024	96 **(50% spinning)**	=Number of drives (12)
Protection:	RAID1, 5, 6, 10	RAID5	RAID1 (N copies)
Max Bandwidth:	6.4GB/s	3.2GB/s (2x ~Xmas'10)	Num drives x 140MB/s = 1.7GB/s
Power:	6,932W (max)	7,445W (max) 2,710W (standby)	1,300W (max)
Tape libraries/drives:	128/1024	8/96	1/12

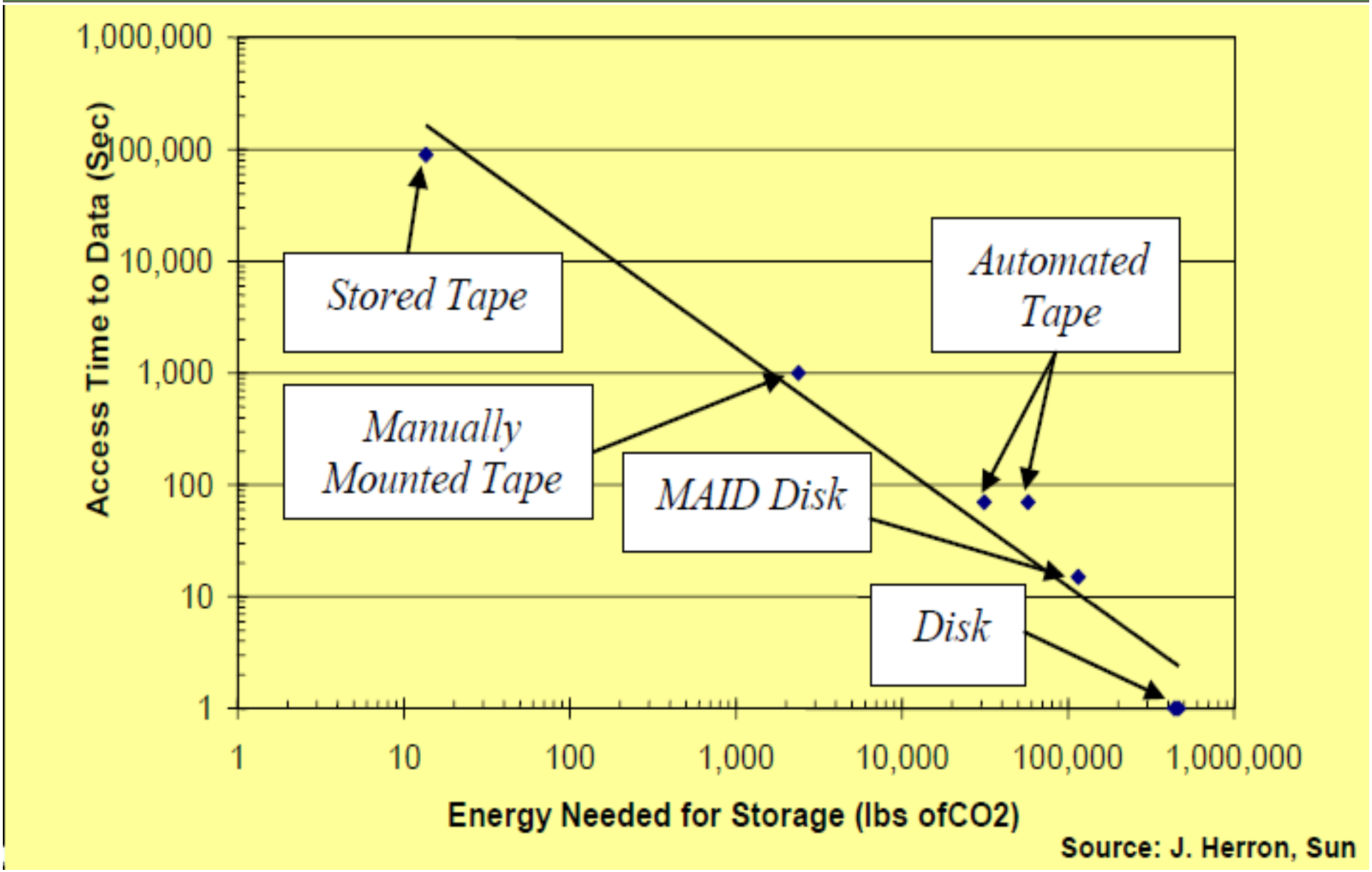
DDN Dynamic MAID

- A tier enabled for D-MAID will go to sleep when the inactivity timer value is reached without any I/O access to the tier
- Drives in sleep mode spin down, but the circuitry is still enabled and they can still accept commands
- User configurable inactivity timer from 5 minutes to 5.5 hours
- For 600 x 2TB SATA drives, normal power usage of 9.24 kW falls to 5.33 kW with D-MAID (57.7%)



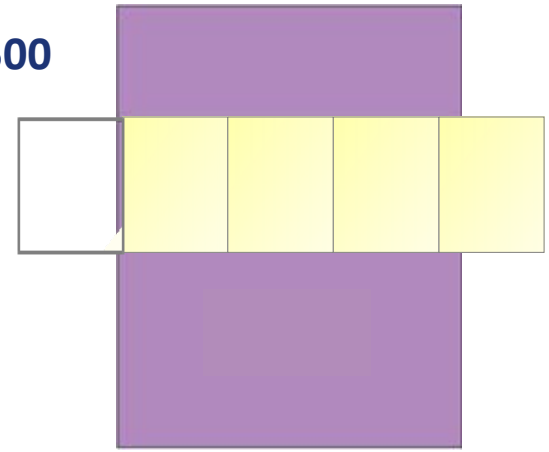
- Maxtor Atlas 15K II current working in Intel IOMeter Random Read pattern.
- Red for the current on +5V power rail (drive electronics)
- Blue for current on 12V power rail (motor and actuator),

How Green?

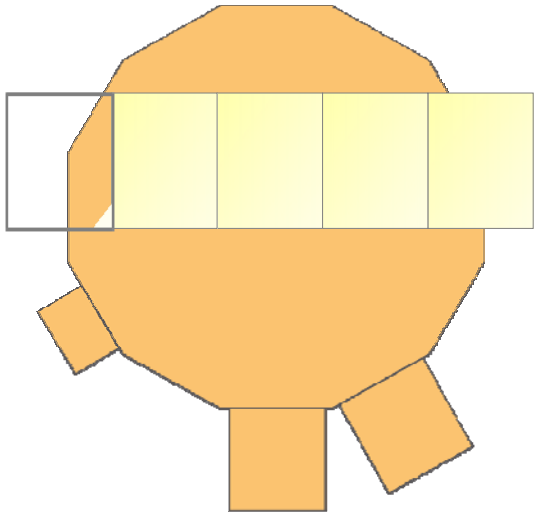

1,000,000
100,000
10,000
1,000
100
10
1
Access Time to Data (Sec)
Stored Tape
Automated Tape
Manually Mounted Tape
MAID Disk
Disk
1
10
100
1,000
10,000
100,000
1,000,000
Energy Needed for Storage (lbs ofCO2)
Source: J. Herron, Sun

Library Footprints

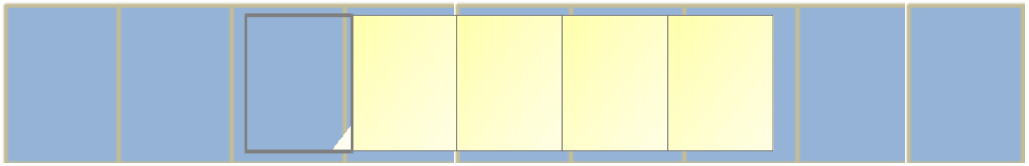
Sun/STK SL8500



Sun/STK 9310 (Silo)

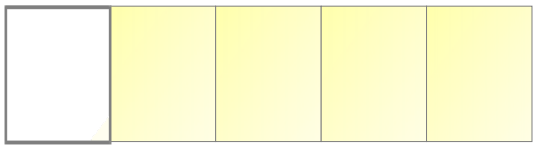


IBM TS3500



Five Frame Spectra T-Finity

With Robotic Service Frames



All configurations shown are floor space requirements (to scale) needed to contain ~3,500 slots and 10 drives.

99c Plastic Magazine

TeraPack™ Architecture:

- 10 LTO cartridges per TeraPack
- Highly efficient use of library wall space
- Utilizes cubic footage of library

T-Finity Scalability

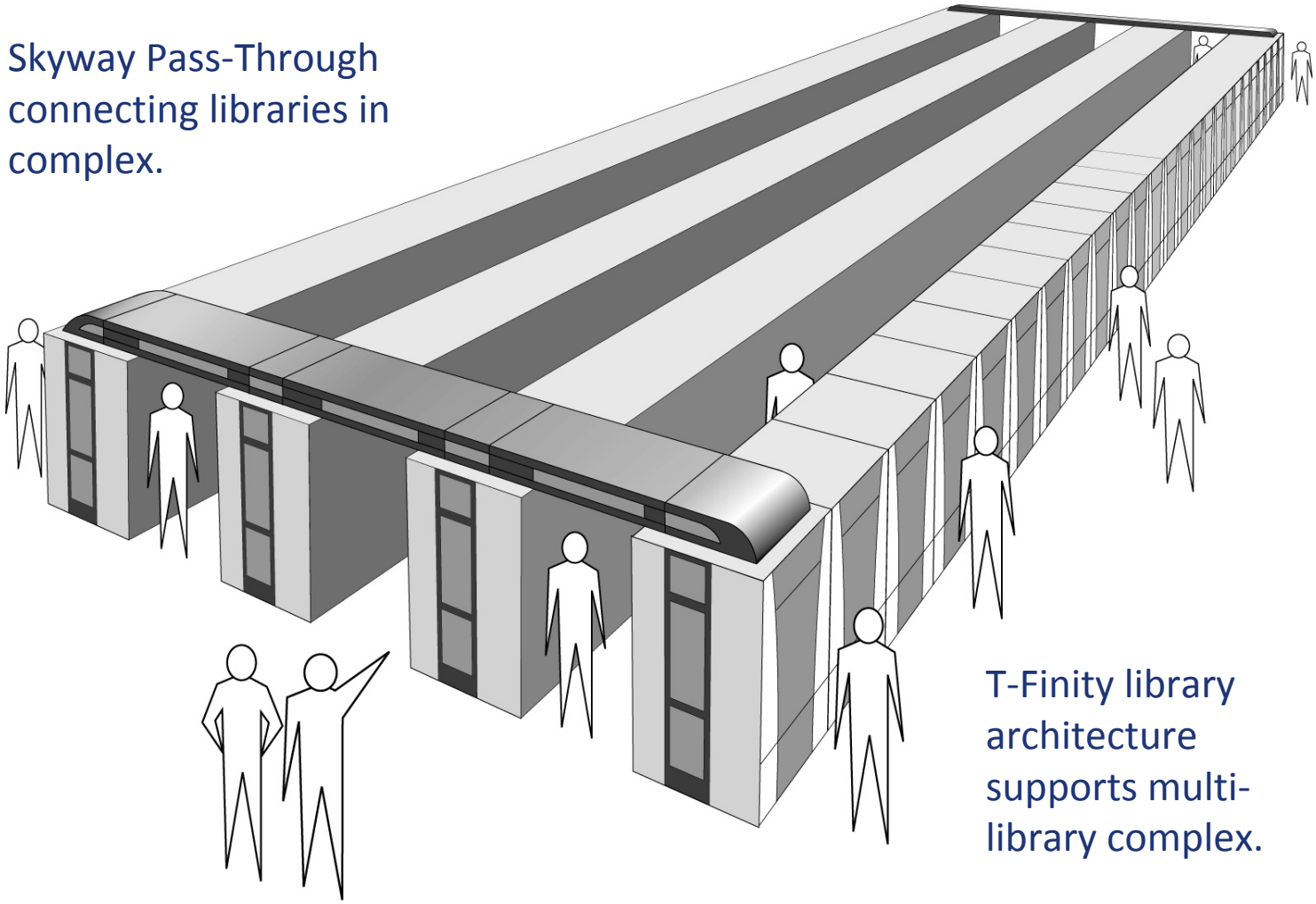
- More frames, slots, and drives than competitors (FC, **SAS, **iSCSI)
- TranScale T200 through T-Finity: Common components, uncommon scalability

Library	Slots	Drives	Frames	Capacity
T-Finity (library complex)	**30,000 (120,000)**	**120 (480)**	**25 (100)**	**45 PB (180 PB)**
SL8500 (library complex)	10,088 (70,616)	64 (448)	6 (42)	15.1 PB (106 PB)
TS3500	6,887	192	16	10.3 PB

** Available with LTO-5

TFinity Complex

Skyway Pass-Through connecting libraries in complex.

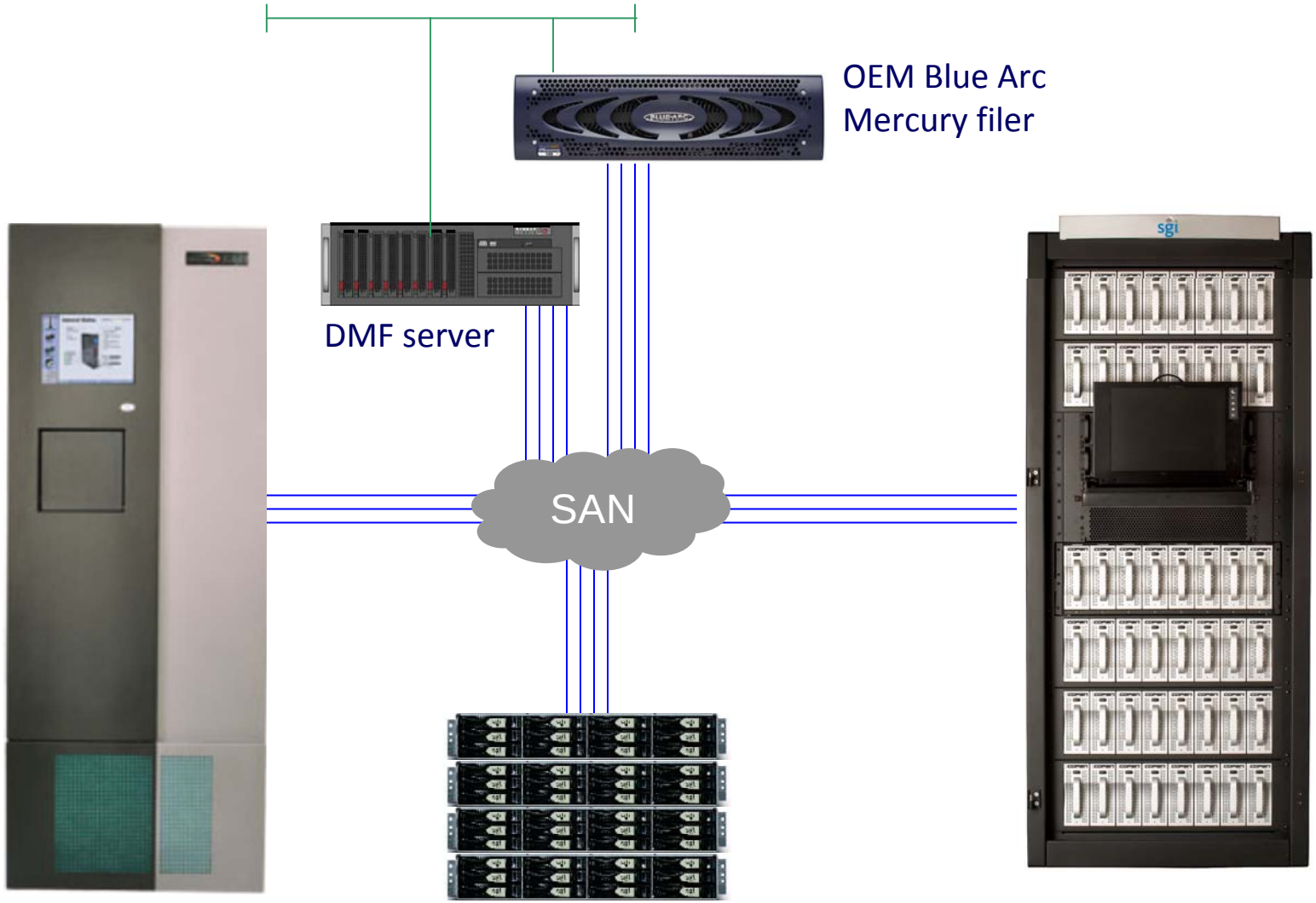


T-Finity library architecture supports multi-library complex.

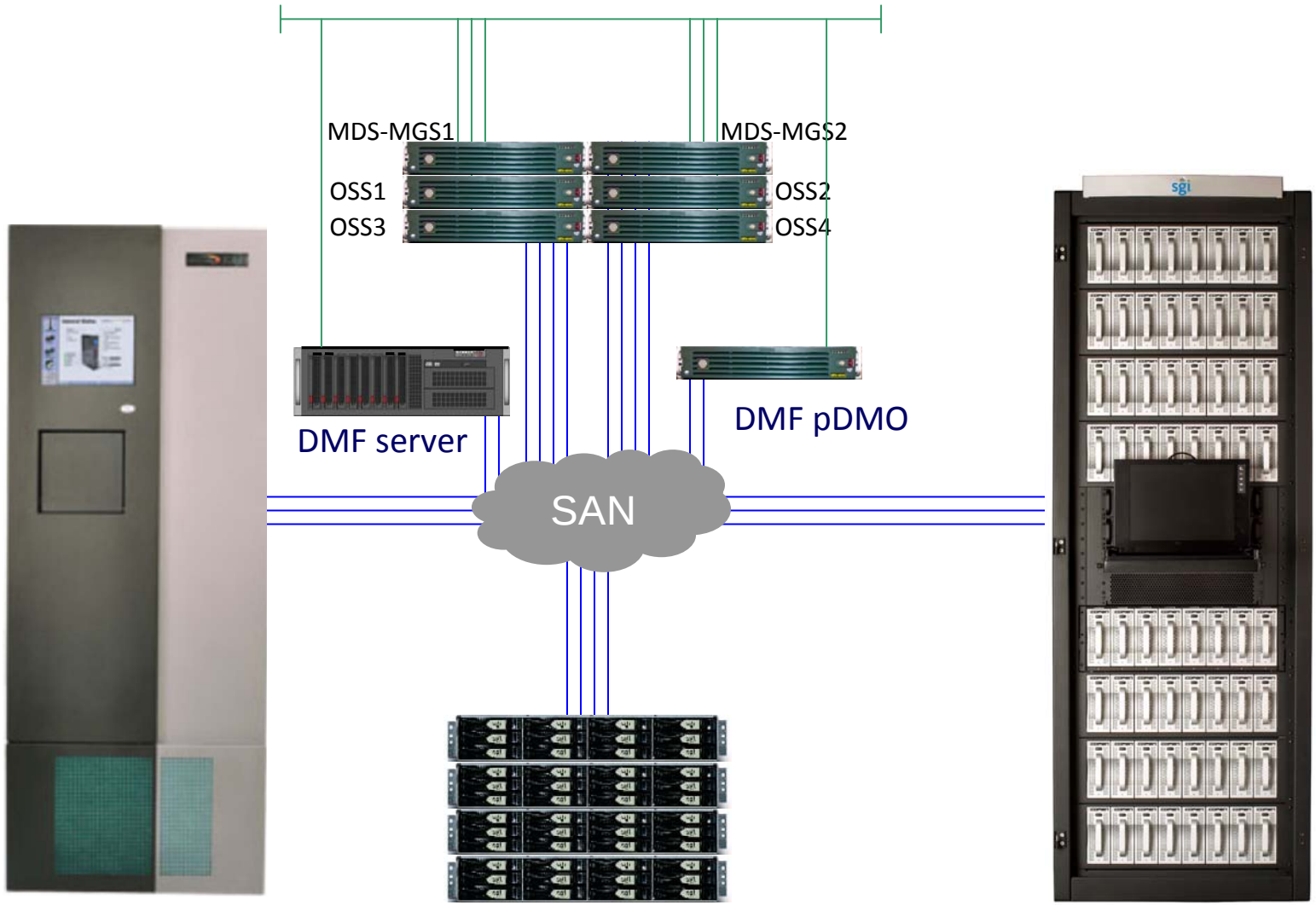
New low end DMF server – ISS3500

- High-density, aggressively-priced storage server
- Single & Dual-socket Intel® Xeon® processors
- High-Availability configuration
  - Redundant 1.4kW Gold Level power supply with PMBus function
  - 7 x 8cm (middle) hot-swap cooling fans
- 4U - 36 hot-swappable 3.5” drives – SAS, SATA & SSD
- Infiniband QDR, GigE & 10 GigE interfaces
- 6x low-profile expansion slots (1 used by MegaRAID card)
- Includes SGI® Storage Management Software and XFS file system
- iSER (iSCSI over InfiniBand), iSCSI, NFS & CIFS
- Ideally suited to Lustre MDS, OSS or DMF server

Active Project – SGI NAS50/100 integration

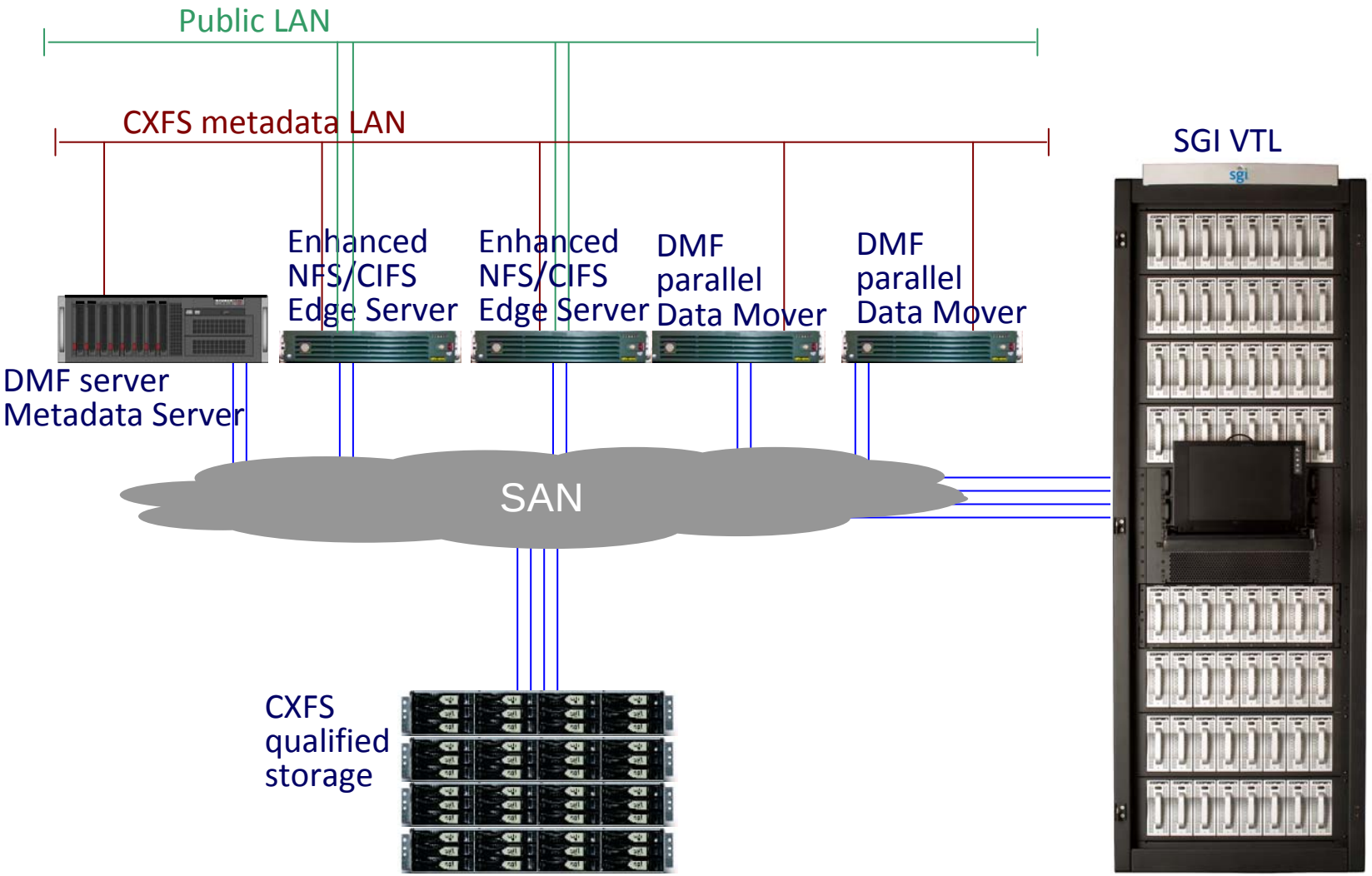


Active Project – Lustre integration



Ready now - Large Scale – Parallel Tape and Edge Serving

Multi tiered solution for incrementally scalable archive and NAS



Future Investment (10PB) Facts

Copan Systems

- Capacity Requirement – 10PB
- Total Annual Power Draw –433.917MW
- Average Cost per KWh - $0.13
- Cost of Power per Year - $56,409.21
- Cost of Power over 5 Years - $282,046.05
- Cooling Ratio - 1.25 kW to cool 1 KW
- Energy Used to Cool - 542.396MW
- Cost to Cool per Year - $70,511.51
- Cost to Cool over 5 years - $352,557.55
- Total Footprint – 110 Square Feet
- Total Copan over 5 years- $634,603.60

Competitor

- Capacity Requirement – 10PB
- Total Annual Power Draw – 2.946GW
- Average Cost per kWh - $0.13
- Cost of Power per Year - $382,980.00
- Cost of Power over 5 Years - $1,914,900.00
- Cooling Ratio - 1.25 KW to cool 1 KW
- Energy Used to Cool - 3.683GW
- Cost to Cool per Year - $478,725.00
- Cost to Cool over 5 Years - $2,393,625.00
- Total Footprint – 390 Square Feet
- Total Initial over 5 years - $4,308,525.00

Total Savings With Copan Solution Over 5 Years

$3,673,921.40

Total Floorspace Savings

280 Square Feet

Future Investment (10PB)

COPAN SYSTEMS

Capacity
10 Petabytes

Footprint
Copan Cabinet Size - 30" x 48"
Total 110 Square Feet

Power & Cooling Cost
Total Anual Power Draw –433.917MW
Averegage Cost per kWh - $0.13
Cost of Power per Year - $56,409.21
Cost of Power over 5 Years - $282,046.05
Cooling Ratio - 1.25 kW to cool 1 KW
Energy Used to Cool - 542.396MW
Cost to Cool per Year - $70,511.51
Cost to Cool over 5 years - $352,557.55
Total Copan over 5 years- $634,603.60

Total Power and Cooling Savings $3,673,921.40

Power & Cooling Cost
Total Annual Power Draw – 2.946GW
Average Cost per kWh - $0.13
Cost of Power per Year - $382,980.00
Cost of Power over 5 Years - $1,914,900.00
Cooling Ratio - 1.25 KW to cool 1 KW
Energy Used to Cool - 3.683GW
Cost to Cool per Year - $478,725.00
Cost to Cool over 5 Years - $2,393,625.00
Total Initial EMC over 5 years - $4,308,525.00

Footprint
Cabinet Size – 24" x 36"
Total 390 Square Feet

Does my archive need more tiers?

- There are many ways to measure the size of an archive

We have more than 1PB managed by DMF

That's a large archive

We have more than 10million files managed by DMF

That's a large archive

We migrate more than 10TB per day to tape

That's a busy archive

We run 120,000 HPC jobs on our HSM FS per week

That's a busy archive

Users complain because their data isn't online

HSM responsiveness can be improved by upgrading a tier or adding a tier