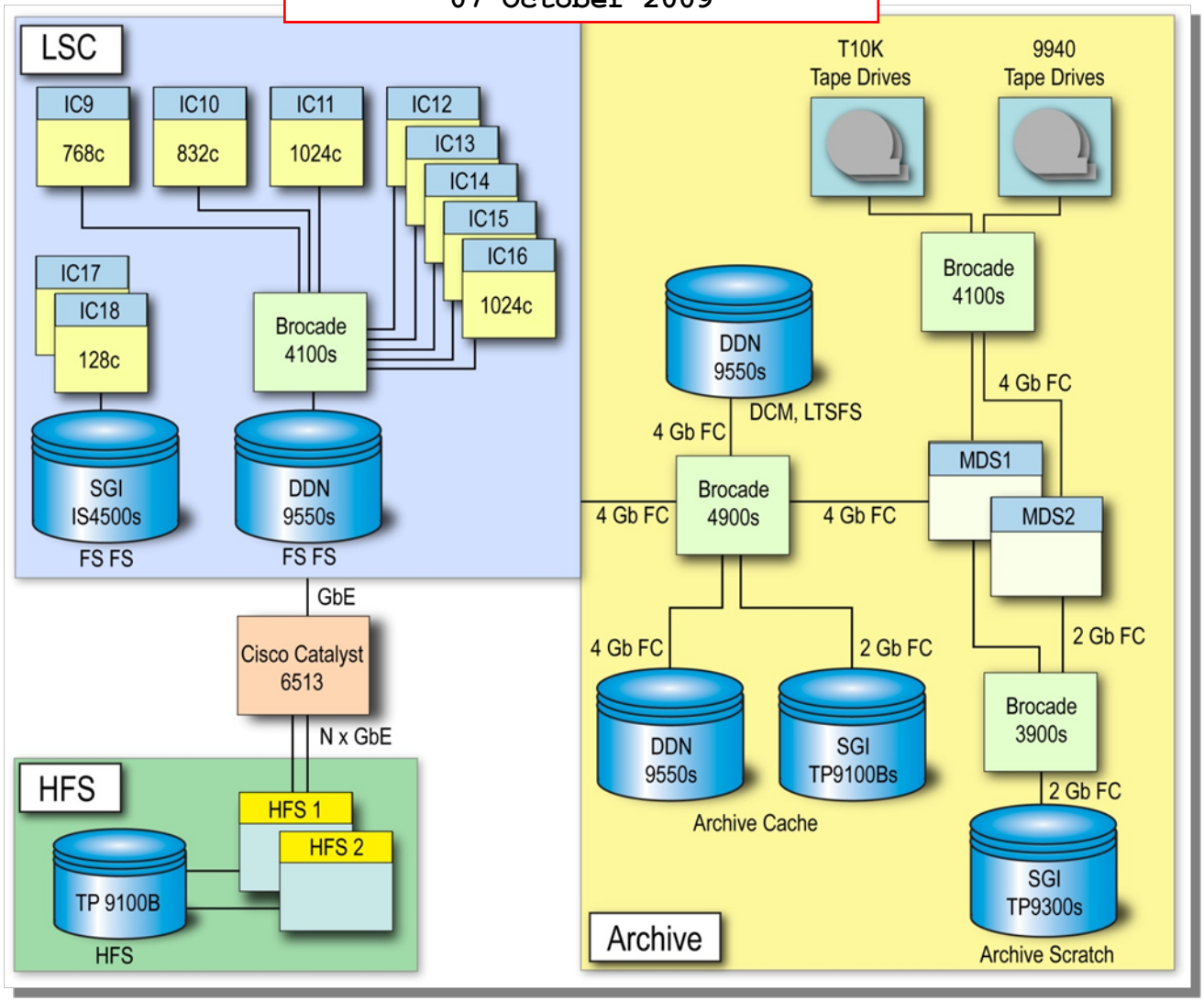
NOAA R&D HPCS RFP
Appendix A.3
~
Current R&D HPCS Architecture
Princeton

07 October 2009



LSC Compute Nodes

All are SGI 4700s (Tollhouses)

- All processors are Itaniums
  - Montecito (IC9, IC10, IC11): 1.6 GHz Clock, 533 MHz Front Side Bus, 24 MB L3 Cache
  - Montvale (IC12 – IC18): 1.67 GHz Clock, 667 MHz Front Side Bus, 24 MB L3 Cache
- Two node board configurations:
  - Single Socket (IC9, IC10, IC11): More memory bandwidth per core
  - Dual Socket (IC12 – IC18): Much less expensive per core
- RAM: 2 GB/core in all systems
- Rack and Stack:
  - Four IRUs per cabinet
  - Ten boards per IRU (Eight CPU boards, two I/O boards)
  - Two NUMALink-4 connections (3.2 GB/s each) per board
  - 128 cores/rack dual socket, 64 cores/rack single socket

LSC Midlife Compute Nodes

Node	Model	Clock Speed	Chip	Node Board	Cores
IC-9	4700	1.6 GHz	Montecito	Single Socket	768
IC-10	4700	1.6 GHz	Montecito	Single Socket	832
IC-11	4700	1.6 GHz	Montecito	Single Socket	1024
IC-12	4700	1.66 GHz	Montvale	Dual Socket	1024
IC-13	4700	1.66 GHz	Montvale	Dual Socket	1024
IC-14	4700	1.66 GHz	Montvale	Dual Socket	1024
IC-15	4700	1.66 GHz	Montvale	Dual Socket	1024
IC-16	4700	1.66 GHz	Montvale	Dual Socket	1024
IC-17	4700	1.66 GHz	Montvale	Dual Socket	128
IC-18	4700	1.66 GHz	Montvale	Dual Socket	128
Total					8000

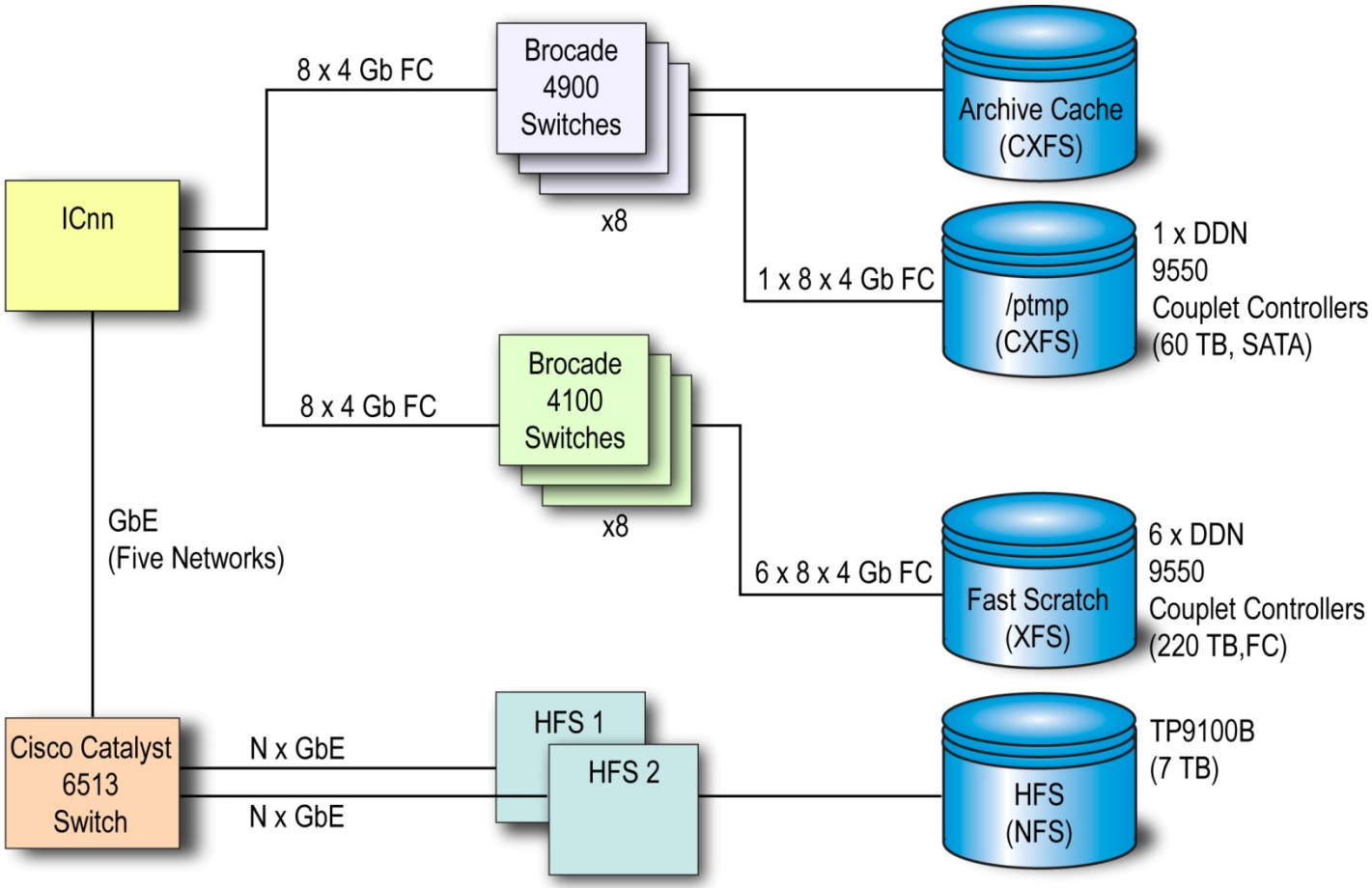
Six 1024 core systems and dedicated post-processing platforms enable more large model runs

Configuration Summary

Subsystem	Princeton Midlife Configuration	Discussion
LSC/IC/AC	IC9: 768c Montecito bandwidth Altix 4700	Provides 8,000 cores and standalone post-processing platforms (the two 128c systems), which isolate the post-processing I/O load from the LSC load, improving system tunability and reducing I/O variability.
	IC10: 832c Montecito bandwidth Altix 4700	
	IC11: 1024c Montecitor bandwidth Altix 4700	
	+ Five 1024c Montvale density Altix 4700	
	+ Two 128c Montvale density Altix 4700	
	8,000 total cores	
Fast Scratch	+ One DDN 9550 couplet controller, with 144 300 GB 10K RPM FC disks and eight 4 Gb FC interfaces. Each array provides 38.4 TB of raw disk space (sixteen LUNs each with eight data disks, or 16 x 8 x 300 GB), a peak throughput of approximately 2.8 GB/s, and peak throughput of approximately 22,000 IO/s.	Provides approximately 124 TB of raw storage with peak throughput of approximately 12.4 GB/s and 502,000 IO/s. This represents a five-fold increase in IO/s performance and is well worth the 20% reduction in total storage space; post-processing demands at Princeton are observed to be transaction-intensive. The additional IO/s will speed up post-processing of the existing workstream and support growing post-processing demand. DDN arrays currently used for post-processing will be used to provide fast scratch for the LSC nodes.
	+ Six LSI IS4500 couplet controllers, each with 64 300 GB 15K RPM FC disks and eight 4 Gb FC interfaces. Each array provides 14.4 TB of raw disk space (twelve LUNs each with four data disks, or 12 x 4 x 300 GB), a peak throughput of approximately 1.6 GB/s, and peak throughput of approximately 80,000 IO/s.	
Home File System	Eight TP9100B enclosures, each with two 6+1+1 LUNs of 73 GB FC drives (existing equipment)	As proposed.
Long Term Scratch File System	Omitted in favor of additional archive cache	Our midlife solution omits this growth in the LTFS in favor of additional archive cache throughput.The LTSFS is not currently fully utilized; Raytheon believes that these resources are better spent to provide additional archive cache throughput.
Archive Metadata Servers (MDSs)	+ Two 96c Madison Altix Bx2 MDSs	Provides Bx2 MDSs as proposed. Provides switches with 100% expansion capability to support additional clients and archive cache devices, protecting the investment in these switches, which typically have a long useful life at Princeton. Provides approximately 38 TB of additional archive cache space and approximately 4.8 GB/s of additional peak archive cache throughput; archive cache throughput is the single most critical resource for scaling the flow of data through the system, and the 4.8 GB/s improvement roughly doubles the current system capability. This solution omits additional DCM space, as proposed, in favor of the archive cache improvement. Retention time within the DCM currently exceeds the period expected when the DCM was fielded, indicating that additional space is not required today; Raytheon believes these resources would be better spent in increasing the archive cache throughput.
Archive Metadata Servers (MDSs)	+ One 8c Madison Altix Bx2 CXFS tiebreaker	
Archive Cache SAN Switching Fabric	+ Nine 32 port Brocade 4900 FC4 switches (expandable to 64 ports each). Ninth switch provided as a spare.	
Tape SAN Switching Fabric	+ One 32 port Brocade 4100 FC4 switch	
Archive Cache	+ Two DDN 9550 couplet controllers, each with 72 300 GB 15K RPM FC disks and eight 4 Gb FC interfaces. Each array provides 19.2 TB of raw disk space (eight LUNs each with eight data disks, or 8 x 8 x 300 GB), a peak throughput of approximately 2.4 GB/s, and peak throughput of approximately 22,000 IO/s.	
DCM	Omitted in favor of additional archive cache	

Storage

Compute Node I/O And Storage



- IC17 and IC18 are direct - attached to the SGI (LSI) IS4500 fast scratch
- IC9 and IC10 have 10 GbE connections to the Catalyst switch

Compute Node I/O And Storage

- Fast Scratch – Large Nodes (IC9 – IC16)
  - XFS
  - Each compute node is attached to the switch fabric (Brocade 4100 switches) with eight 4 Gb FC ports
  - Six DDN 9550 couplet controllers serve the Fast Scratch file systems
    - Each couplet controller is connected to the switch fabric with eight 4 Gb FC ports
    - A couplet controller with 16 tiers of drives is capable of sustaining 2.4 GB/s and up to 22,000 file system IO operations per second (using 4 KB blocks)
    - Each couplet controller has 4 GB cache (net usable)
  - Fast Scratch has 92 tiers of 300 GB FC drives (approximately 2.4 TB per 8+1 tier)
  - 1024c nodes (IC11 – IC16) are allocated 12 tiers each (~29 TB, ~1.8 GB/s sustained)
  - IC9 and IC10 are allocated 8 tiers each (~19 TB, ~1.2 GB/s sustained)

Compute Node I/O And Storage

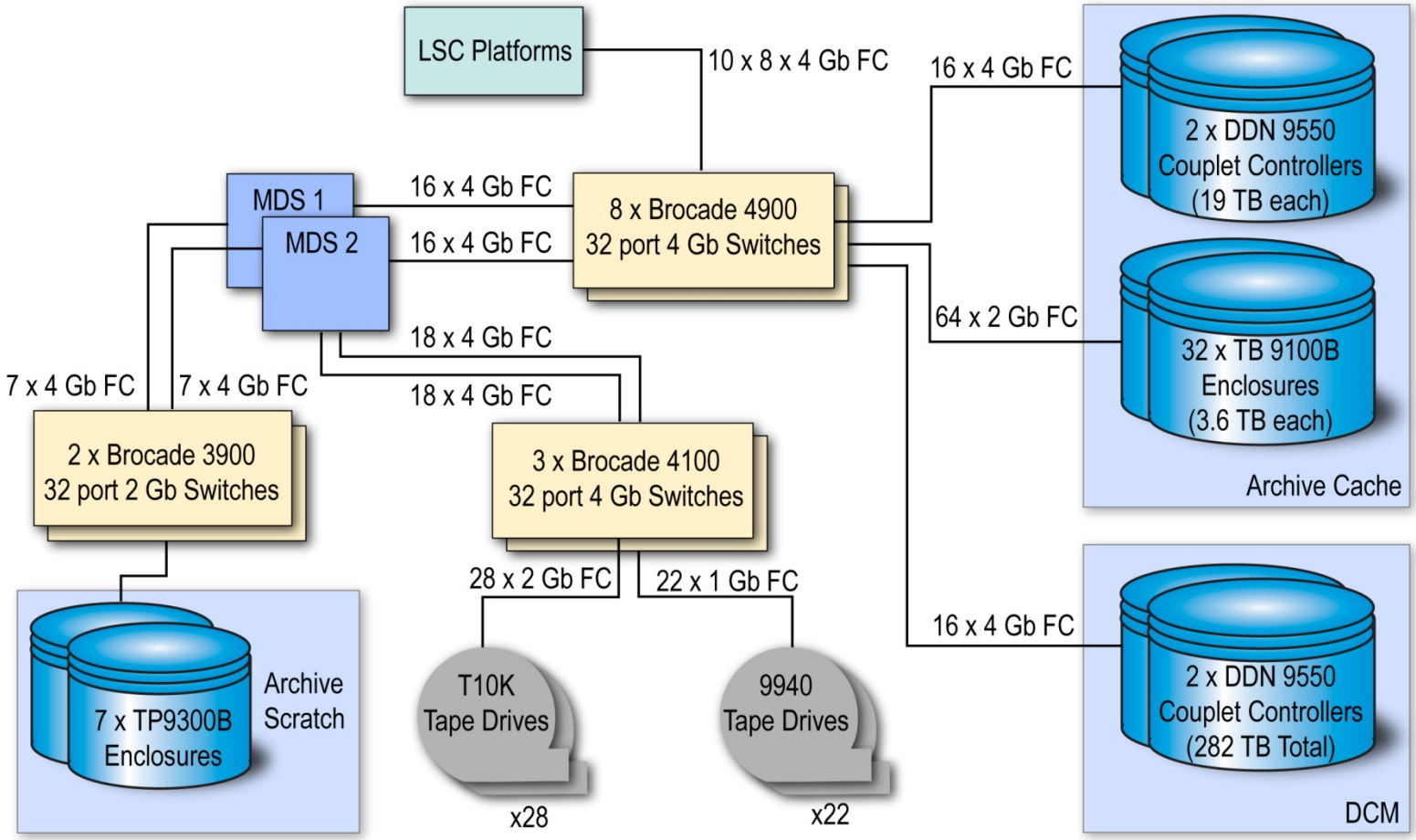
- Fast Scratch – Post-Processing (IC17 and IC18)
  - XFS
  - Each post-processing compute node is direct attached to the IS4500 arrays with twelve 4 Gb FC ports
  - Six SGI/LSI IS4500 couplet controllers serve the post-processing Fast Scratch file systems
    - A couplet controller with 64 drives is capable of sustaining 1.4 GB/s while reading, 1.0 GB/s while writing, and up to 80,000 file system IO operations per second (using 4 KB blocks)
    - Each couplet controller has 2 GB cache
    - A total of 72 (4+1) LUNs of 300 GB 15K RPM FC drives are provided (1.2 TB per LUN, 43 TB per node, 86 TB total)

Compute Node I/O And Storage

- Home File System (HFS)
  - All systems are connected to the HFS via NFS over gigabit ethernet (GbE)
  - The HFS capacity is expanded to 7 TB
  - The network connectivity of the HFS servers (HFS1 and HFS2) will be expanded to 2 x GbE or 4 x GbE (TBD) via trunking/bonded etherchanneling
  - A costed option is offered to expand the HFS by filling the remainder of the existing SGI SATA array with eight tiers (8 + 2) of 1 TB SATA drives, adding approximately 64 TB of usable space
- Long Term Scratch File System (LTSFS)
  - CXFS
  - All nodes (IC9 – IC18) connected to LTSFS (/ptmp) via archive FC connections (eight 4 Gb connections)
  - LTSFS configuration is unchanged at midlife

Archive

Archive Configuration



Archive Configuration

- New Metadata Servers
  - Two new metadata servers, MDS1 and MDS2, will replace AC1 and AC2. Each will have:
    - 96 BX2 processors (1.5 GHz Madison 9M processors)
    - 192 GB RAM
    - Sixteen 4 Gb FC ports to access the archive cache and DCM
    - Fourteen 4 Gb FC ports to access the Titanium drives
    - Four 4 Gb FC ports to access 9940 tape drives
    - Seven 4 Gb FC ports to access archive scratch
  - A new 8c Altix voting server will be provided to support CXFS

Archive Configuration

- Archive Cache
  - Archive cache is expanded with the addition of two DDN 9550 couplet controllers and 16 tiers of 300 GB 15K RPM FC drives:
    - Total archive cache storage increases from 115 TB to 153 TB
    - Total archive cache throughput increases from approximately 3,840 MB/s to approximately 7,840 MB/s
  - The archive cache SAN is provided by eight Brocade 4900 switches
    - 32 ports each, expandable to 64 ports each
    - 4 Gb
    - Ninth switch provided as spare
- DCM
  - DCM is unchanged at midlife

Archive Configuration

- Archive Scratch
  - Archive scratch will re-use seven TP9300 enclosures (formerly /vftmp devices):
    - Total archive scratch storage increases to 6,132 TB
    - Total archive scratch throughput increases to approximately 1,400 MB/s
  - The archive scratch is switched and fenced to support MDS failover by re-using two existing Brocade 3900 switches
  - Improved performance of the archive scratch is critical for
    - Faster back-ups, allowing greater data assurance and enabling greater operational flexibility
    - Faster data migration

Archive Configuration

- Archive Tape Resources
  - 28 T10K tape drives connected to tape SAN (using a single port connection from the tape drives to provide better stability)
  - 22 9940 tape drives connected to tape SAN
  - Tape SAN fabric provided by three 32 port, 4 GB Brocade 4100 switches
  - 9840 tape drives retired before MDS replacement
  - 9940 tape drives supported until all 9940 tapes migrated to Titanium (approximately 15 months)
  - Brocade 2800 switches (old tape SAN fabric) retired