



Non-shared disk cluster - a fault tolerant, commodity approach to hi-bandwidth data analysis.

Doug Olson,
E. Hjort, J. Lauret, M. Messer,
E. Otoo, A. Shoshani, A. Sim

CHEP'01
3-7 Sept. 2001, Beijing





Outline

- Motivation, shared vs. non-shared
- Current status
- Planned implementation
- Schedule



5 Sept. 2001

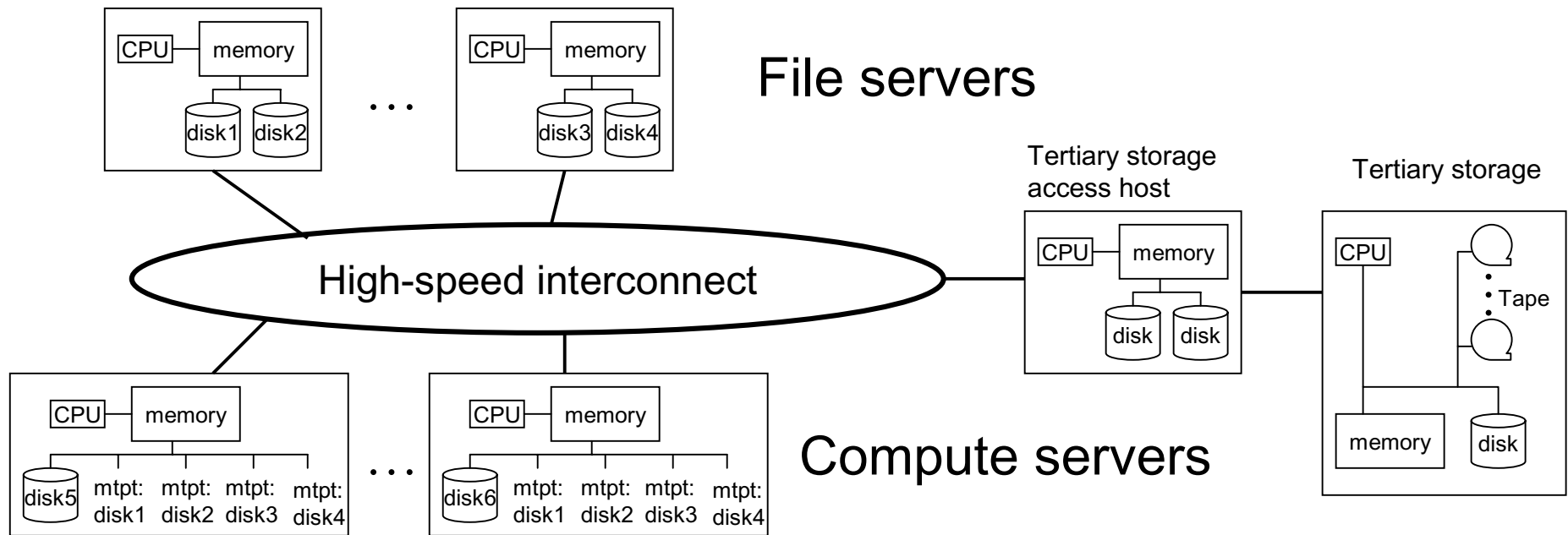
D. Olson, SN-cluster, CHEP01

2





Example network-shared disk configuration





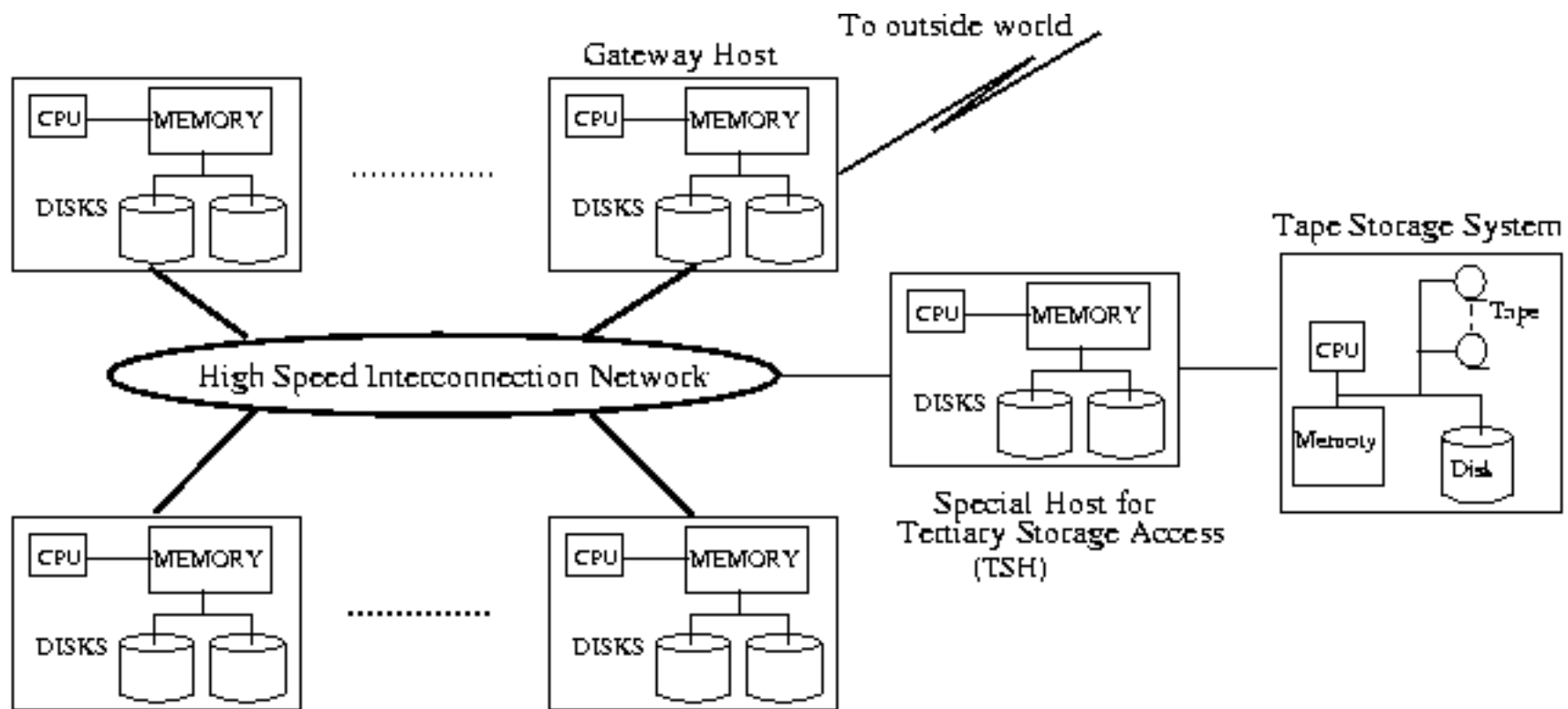
Network-shared disk pros/cons

- Advantages
 - Simpler system & data management
- Disadvantages
 - Single points of failure
 - Extra expense in making these robust, and this is not always successful
 - Shared disk-drive access by many processes reduces bandwidth significantly compared to streaming mode





Example configuration of "shared-nothing" cluster



5 Sept. 2001

D. Olson, SN-cluster, CHEP01

5





"Shared-nothing" cluster pros/cons

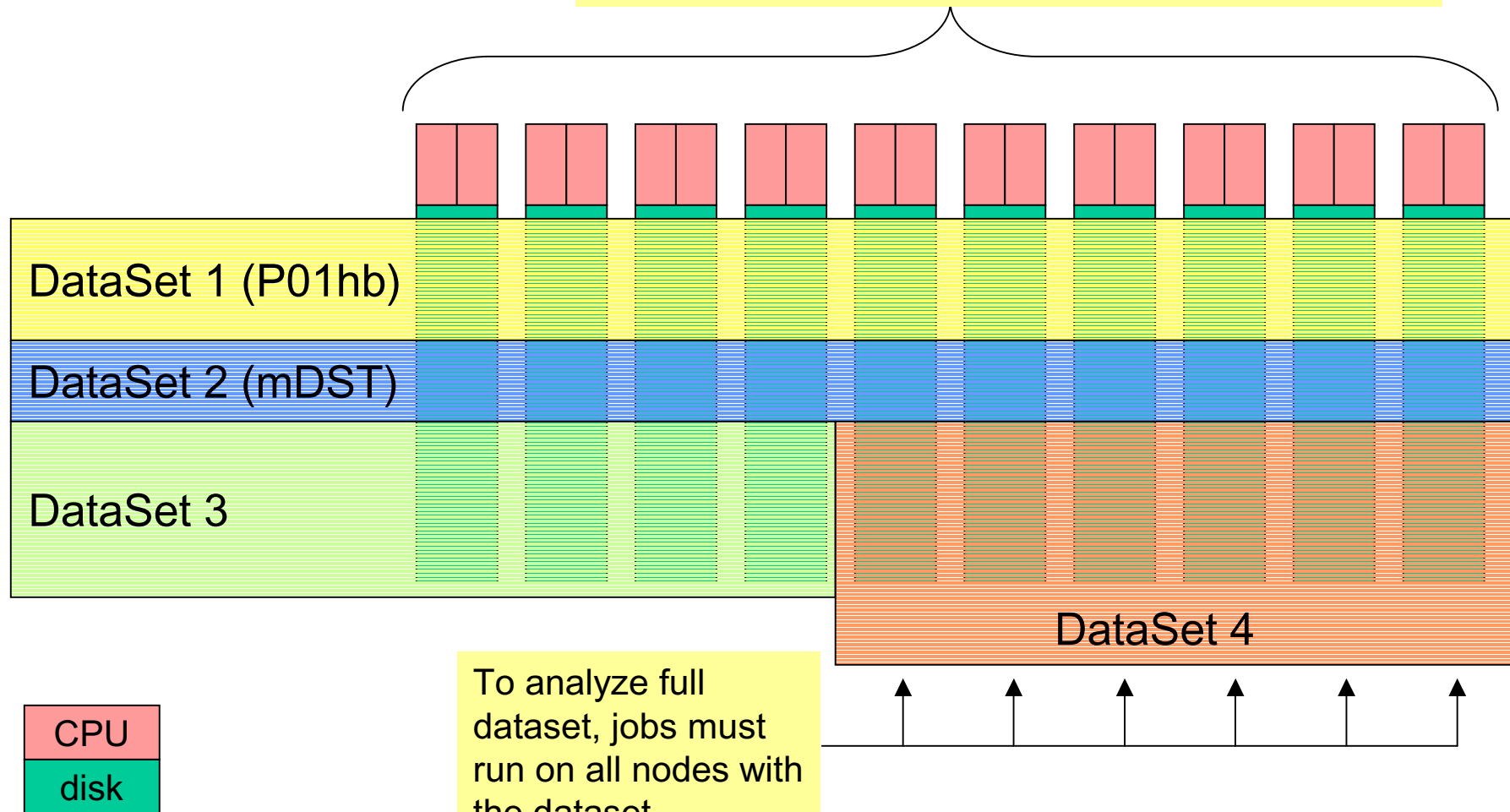
- Advantages
 - Inexpensive hardware
 - Higher bandwidth to each disk due to reduced number of processes accessing a disk simultaneously
 - The more complex processing & data management software should have load balancing features for both CPU and I/O load
- Disadvantages
 - Reduced reliability of components requires an approach that tolerates failures
 - More complex processing & data management





Current status

A dataset is "striped" across some or all nodes



5 Sept. 2001

D. Olson, SN-cluster, CHEP01

7





High Bandwidth Nodes at PDSF - Microsoft Internet Explorer

Address <http://www-rnc.lbl.gov/PDSF/HB.html>

Usage: HBjobs.pl [options] <userscript> <filelist>

Options:

- q: specify queue (defaults to short).
- s: submit jobs to LSF queue.
- v: verbose mode, prints details.
- d: provide a string which will be required to be part of the dataset name.
- u: provide an email address for LSF job completion emails.
- j: provide an identifier to be combined with filenames to create jobnames.

Arguments:

<userscript> is the script that will be executed once for each file in <filelist>. It should take two arguments: First, a directory path to the input file, and second, the name of the input file. The directory path is determined automatically by HBjobs.pl and the name of the input file comes directly from <filelist>. There are examples in ~starofl/highbandwidth, e.g., testjob.csh.

<filelist> is a list of files you want to analyze. For examples look ~starofl/highbandwidth/lists. To analyze all files in a dataset just specify <dataset>.masterfilelist.txt as your filelist. To analyze a subset of a dataset make a copy of <dataset>.masterfilelist.txt and edit the copy as needed. Do not edit the library files themselves. Combining datasets into one file list is OK subject to use of the -d option.

File lists: In the directory ~starofl/highbandwidth/lists there is a set of files for each dataset:

Filename	Description
<dataset>.masterfilelist.txt	All files in that dataset
<dataset>_<partition#>_<#partitions>.txt	Files in each partition
<dataset>.nodelist.txt	List of nodes where files are
<dataset>.nodemap.txt	Mapping between nodes and partitions
<dataset>.dirfile.txt	path on HPSS and path on local disk

Done Internet





High Bandwidth Nodes at PDSF - Microsoft Internet Explorer

Address <http://www-rnc.lbl.gov/PDSF/HB.html>

Dataset	Description	Total size (GB)	Total # Files
P00hm_central_2000_08	DST files	151	1400 (x4)
P00hm_central_2000_09	DST files	94	839 (x4)
P00hm_minbias_2000_08	DST files	76	1432 (x4)
P00hm_minbias_2000_09	DST files	41	730 (x4)
P01he_central_2000_08	DST files	387	3532 (x4)
P01he_central_2000_09	DST files	138	1261 (x4)
P01he_minbias_2000_08	DST files	65	1176 (x4)
P01he_minbias_2000_09	DST files	35	621 (x4)
P01hf_minbias_2001_202	DST files	5.5	34 (x4)
P01hf_minbias_2001_203	DST files	14	84 (x4)
P01hf_minbias_2001_204	DST files	17	140 (x4)
P01hf_minbias_2001_206	DST files	10	84 (x4)
P01hf_minbias_2001_209	DST files	32	247 (x4)
P01hf_minbias_2001_210	DST files	52	386 (x4)
P01hf_minbias_2001_212	DST files	28	224 (x4)
P01hf_minbias_2001_213	DST files	34	165 (x4)
P01hf_minbias_2001_214	DST files	48	192 (x4)
P01hf_minbias_2001_217	DST files	3.7	67 (x4)
P01hf_minbias_2001_218	DST files	38	372 (x4)
P01hf_minbias_2001_220	DST files	80	622 (x4)
P01hf_minbias_2001_227	DST files	59	656 (x4)

Done Internet



5 Sept. 2001

D. Olson, SN-cluster, CHEP01

9





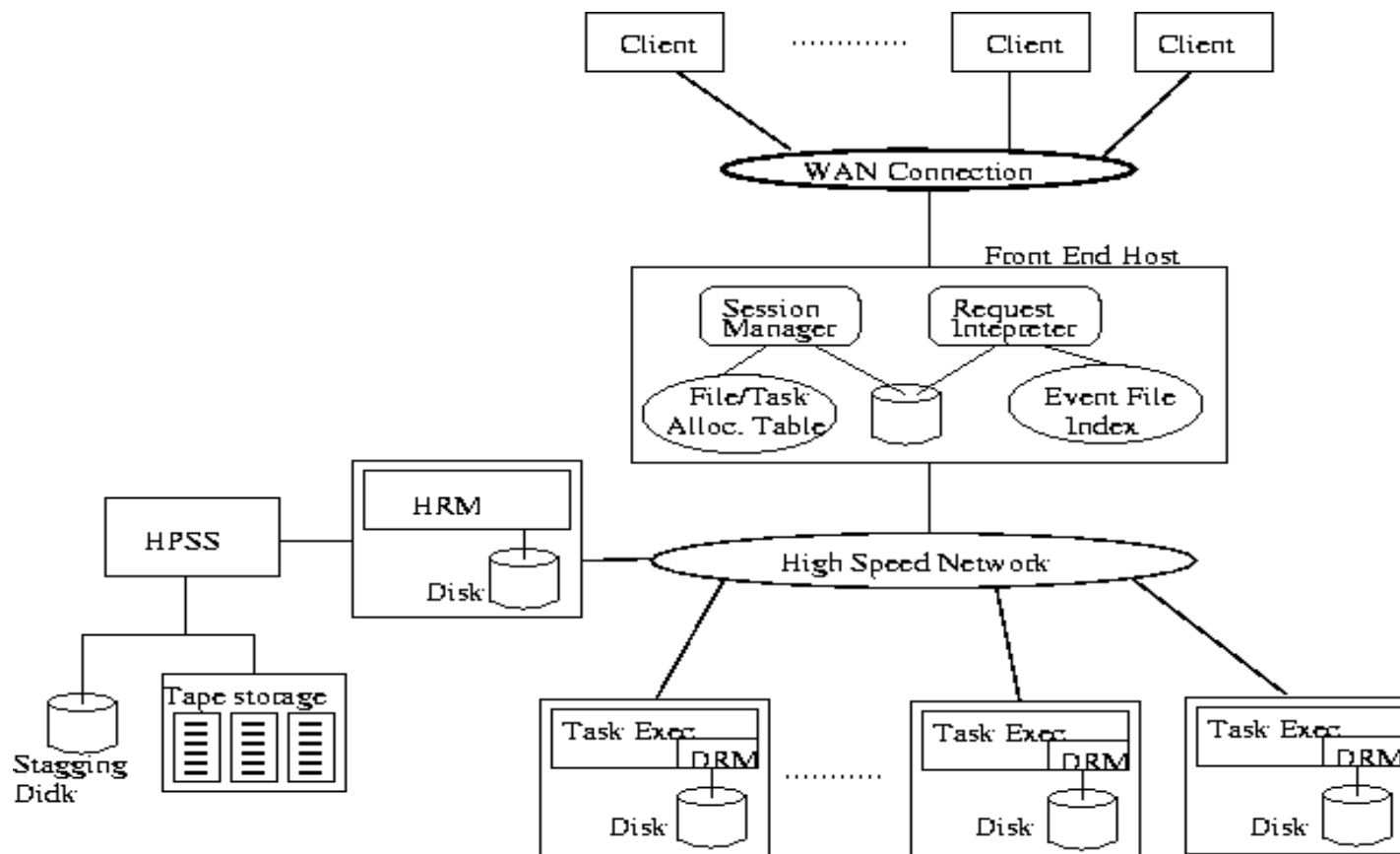
Planned implementation

- Current versions of software components will be extended to meet SN-cluster needs
 - HRM, DRM (Hierarchical & Disk resource managers)
 - Event-level bit-sliced index
- Develop a Session Manager
 - interfaces with the cluster management software such as LSF or Condor-G
- Develop and adjust policies for load balancing that includes automatic data replication





Module Configuration for Data Management in SN-Cluster Computing



5 Sept. 2001

D. Olson, SN-cluster, CHEP01

11





Implementation Phases

- Phase 1: Develop strategies for dynamic file allocation and load balancing the task execution within the hosts to obtain optimal file processing schedules.
- Phase 2: Implement a prototype system that includes a Session Manager (SM) and the storage resource managers ccDRM and cCHRM that work cooperatively with cluster management software (CMS) to achieve optimal scheduling policies of tasks.
- Phase 3: Develop and implement a protocols for the fail-safe and recoverable execution of tasks as specialized distributed transactions on an SN-cluster of workstations.

