Data Intensive High Performance Computing

Challenges for the Future

Marc Snir
Faiman-Muroga Professor
Dept Head

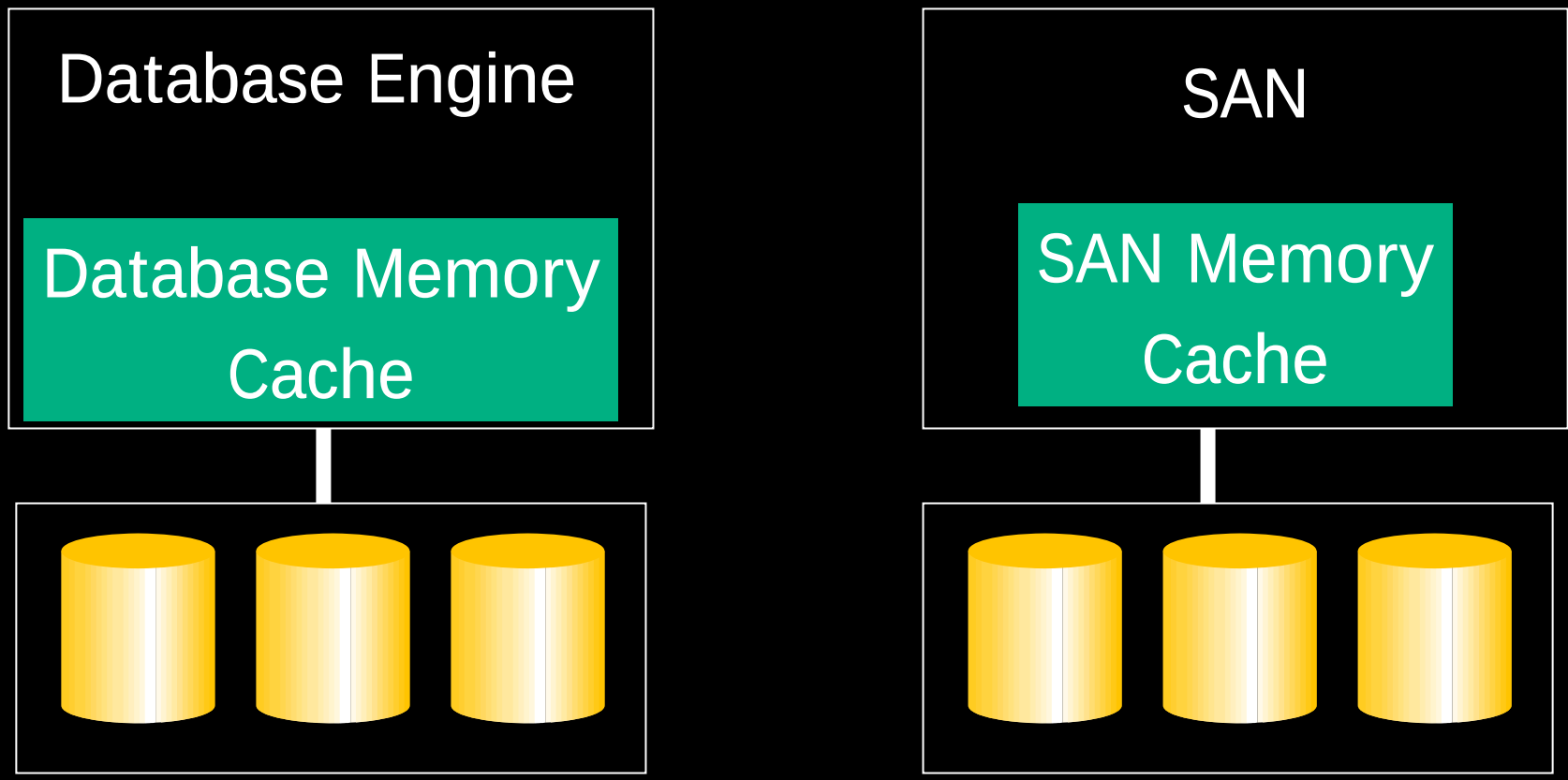
Supercomputing Aurea Mediocritas

- US govt. has lost faith in the “submarine model” for supercomputer development
  - “capability” has shifted from “custom” to “some custom components” (~~custom~~/hybrid/commodity)
- Game is now defined as follows:
  - Find technologies that are being developed for a mass market and can be reused in HPC
  - Find minimum change needed to adapt to HPC
- Mass storage and large-scale data management has a large market
- Correlated access to mass storage by many processors is an HPC need
  - space correlated
  - time correlated

Breaking Space and Time Correlation

- Space coordination: supporting efficiently fixed permutations from application distribution to storage distribution any permutation
  - Any permutation can be expressed as a composition of two “random” permutations (Valiant, Upfal).
  - Can be supported efficiently on multistage networks
- Time coordination: Smoothing “impulse” I/O into continuous I/O
  - Time domain equivalent of space domain scheme above
- Both transformations can be used to improve disk performance (increased block size, reduced seeks)
- What is the right structure for an I/O “space-time randomizer”?

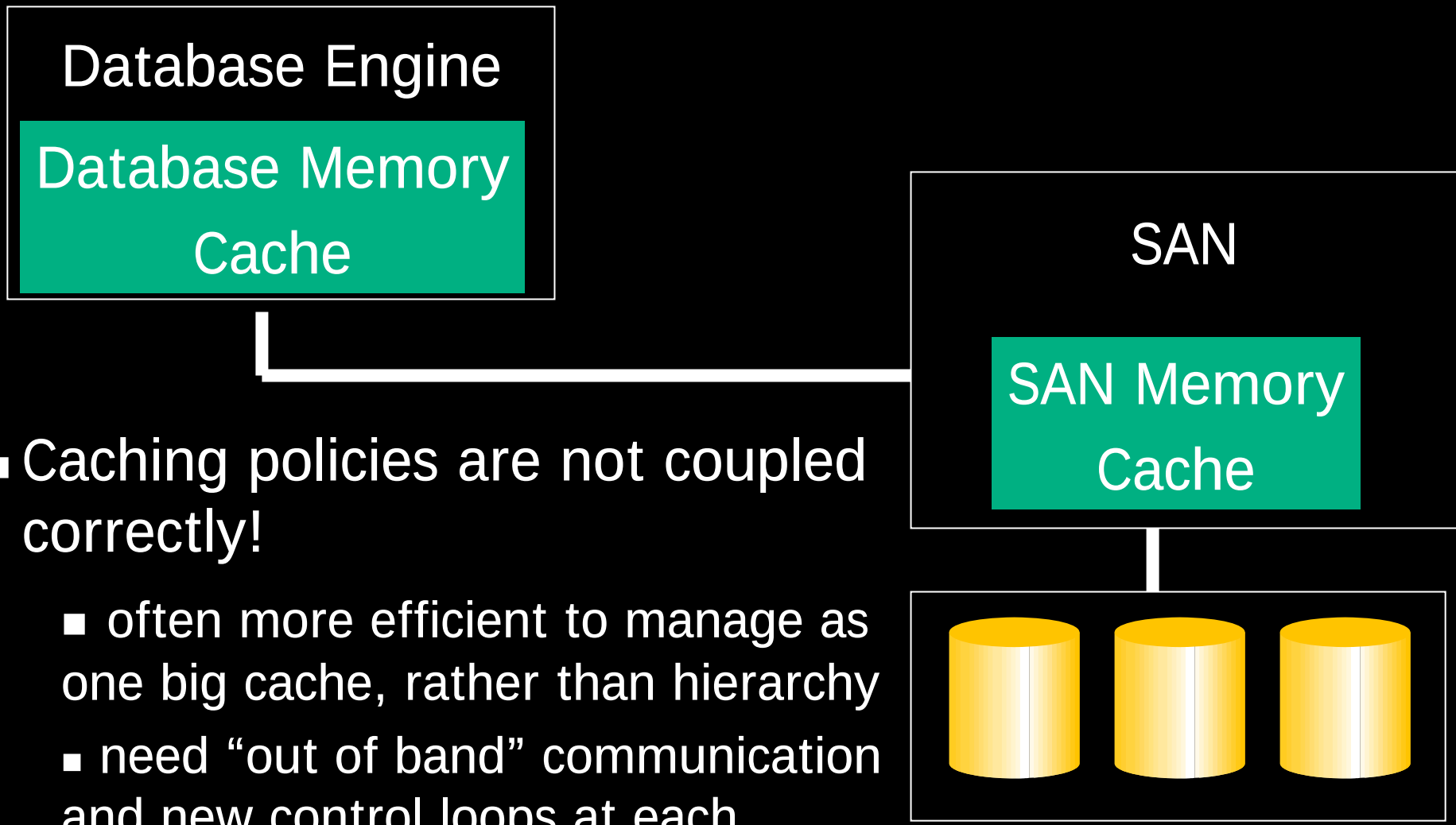
Control Loops



- Cache management: control loop that adjusts cache configuration based on inputs from local sensors

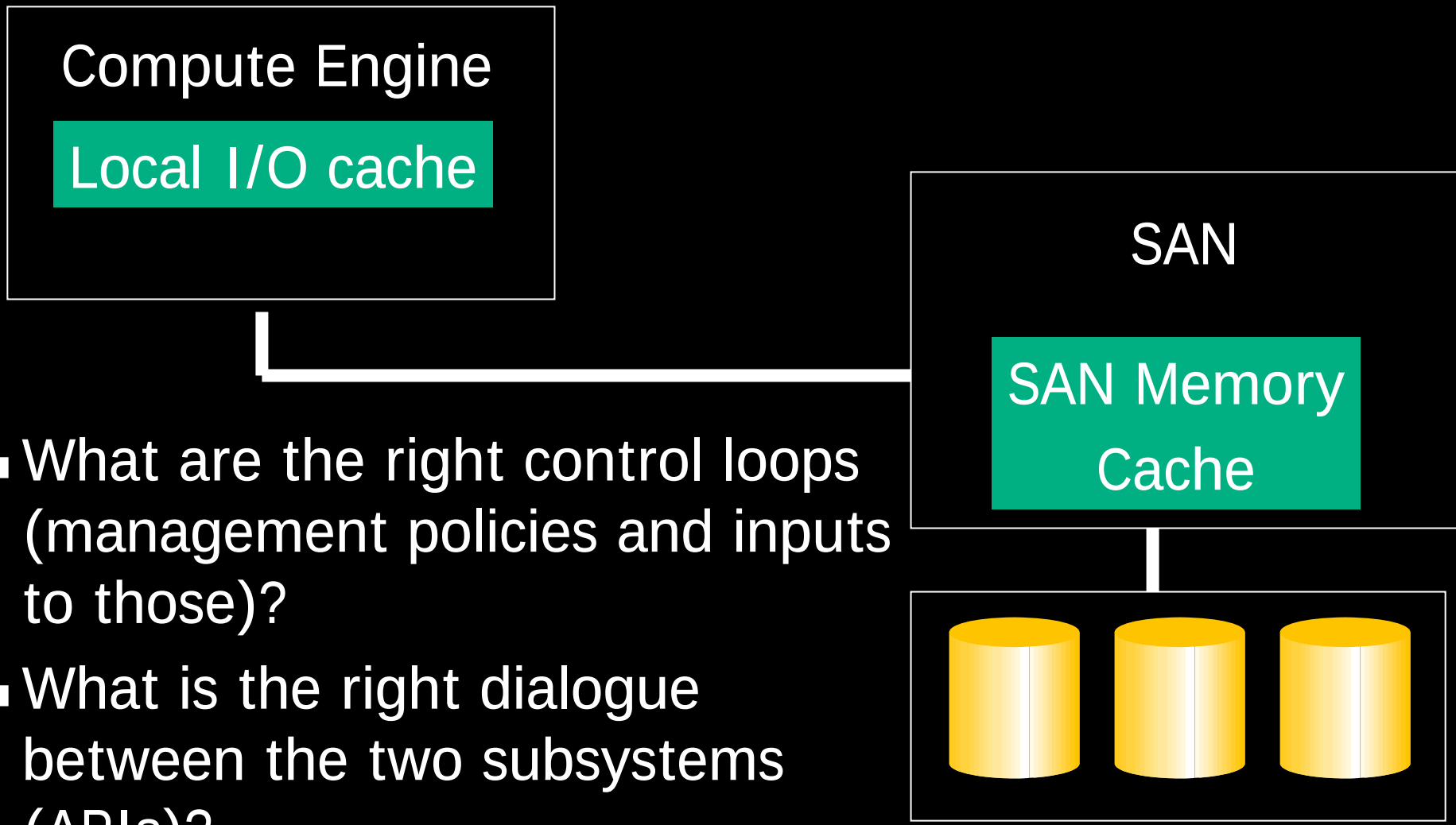
YY Zhou

Coupled Control Loops



- Caching policies are not coupled correctly!
  - often more efficient to manage as one big cache, rather than hierarchy
  - need “out of band” communication and new control loops at each system

Coupled control loops (continued)



- What are the right control loops (management policies and inputs to those)?
- What is the right dialogue between the two subsystems (APIs)?

Self Inflicted Problems

- Increase in number of metadata updates (32K file creates per second)
  - Number of files should be proportional to number of jobs, not number of processes
  - Parallel computations handle concurrency control in the algorithm – they don't assume a globally coherent, cached memory; same should apply to I/O space.
  - Coupling of distinct jobs through files often has a simple structure – e.g. pipeline; should optimize to this

Architecture Research: Memory Hierarchy

Reality

	Register	L1	L2	Memory	Disk
Access time	1	3	20	150	$2x10^8$
Size	100's	64KB	1M	1GB	1TB

Ideal

	L0	L1	L2	L3	L4	L5
Linear	1	3	20	150	230	300
Exp	100's	64KB	1M	1GB	0.3TB	10TB

Will consumer market push for "solid state disks" provide new architectural opportunities?

Programming Models

- Alternatives exist: OS/400 (iSeries) – flat address space
  - do they make sense given the huge latency gap?
  - do file systems and hierarchical directories make sense?

Longhorn (2003 view)

- The successor to Windows XP (due in 2004, and rapidly slipping to 2005) is currently code named Longhorn...
- The most important feature of Longhorn is replacement of the familiar DOS/Windows filesystem with an object database...
- The Longhorn filesystem will be based on the technology of a re-thought and expanded SQL Server database (the project coded Yukon)...

Longhorn (2005 view)

- One thing that has changed is that the initial release of Longhorn will no longer include the Windows Future Storage (WinFS) relational database-based storage engine as originally planned...
- WinFS is implemented as an add-on to NTFS and is not a completely new file system. Rather, it is a new storage engine built on the NTFS file system...