

HPSS at Los Alamos

Gary Lee

Mail Stop B269

Data Storage Systems (CIC-11) Group

Los Alamos National Laboratory

Phone:+1-505-667-2828; FAX:+1-505-667-0168

email: rgl@lanl.gov

URL: <http://storage.lanl.gov/cic11/hpss.html>

Presented at the THIC Meeting in Albuquerque NM

April 21, 1998

Outline

- Overview of the High Performance Storage System (HPSS)
- Current Status at Los Alamos
- Accelerated Strategic Computing Initiative (ASCI) Requirements
- Challenges
- Vision

Overview of HPSS

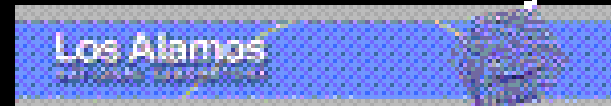
- Scalable, parallel, high-performance software system
- A collaborative effort
- Vendor supported - IBM
- A major ASCI project
- Winner of a 1997 R&D 100 Award

HPSS Collaborators

Lawrence Livermore
National Laboratory



Los Alamos
National Laboratory



Sandia
National Laboratories



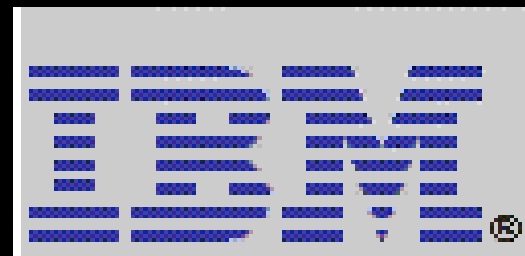
Oak Ridge
National Laboratory



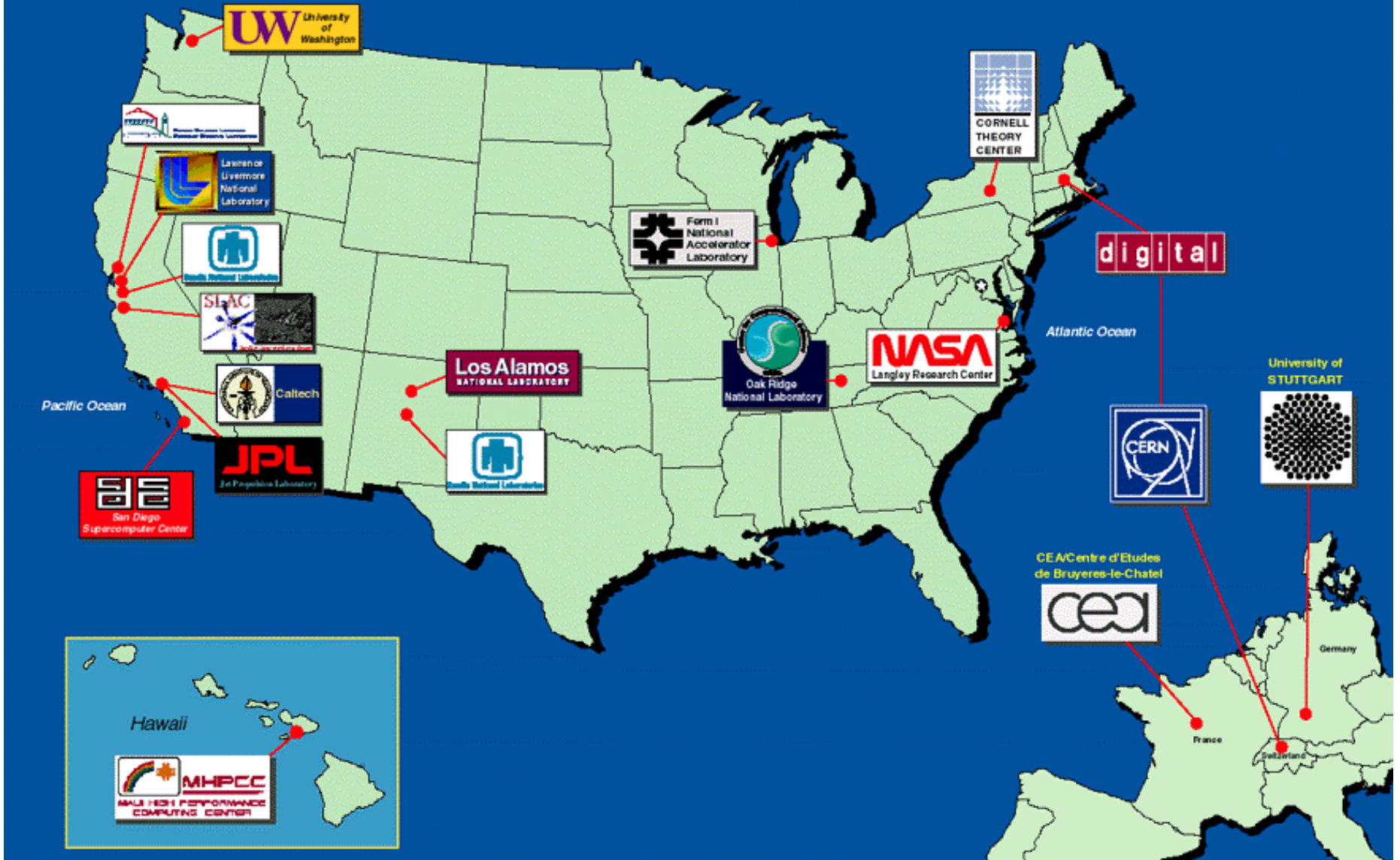
Lawrence Berkeley
National Laboratory



IBM



HPSS Deployment Partners

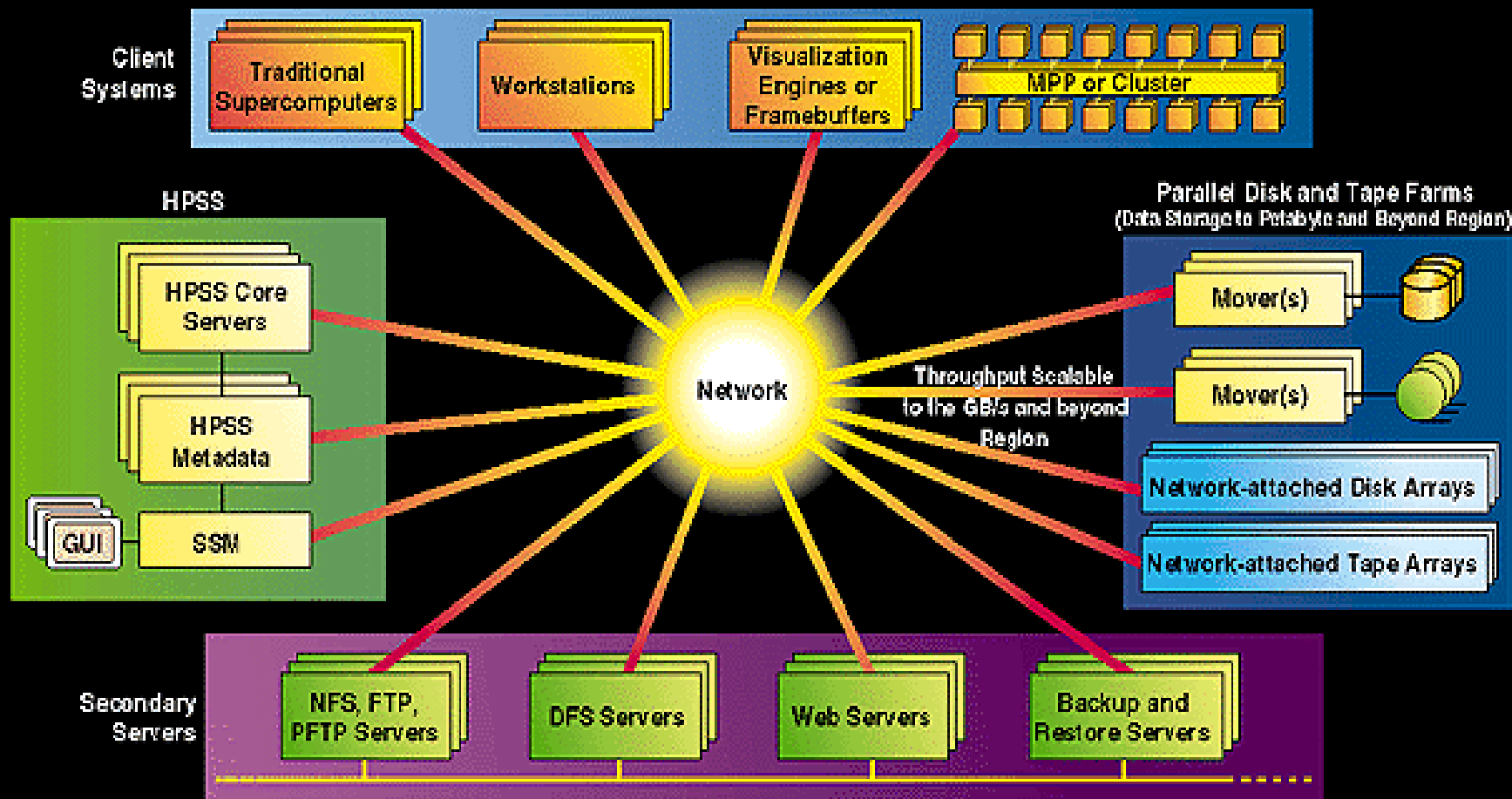


Overview of HPSS

- High Capacity
 - store petabytes of data and billions of files
- High Performance
 - data transfer rates in the GB/sec range
- Parallel data transfers across disks, tapes, and networks

System Architecture Supported by HPSS

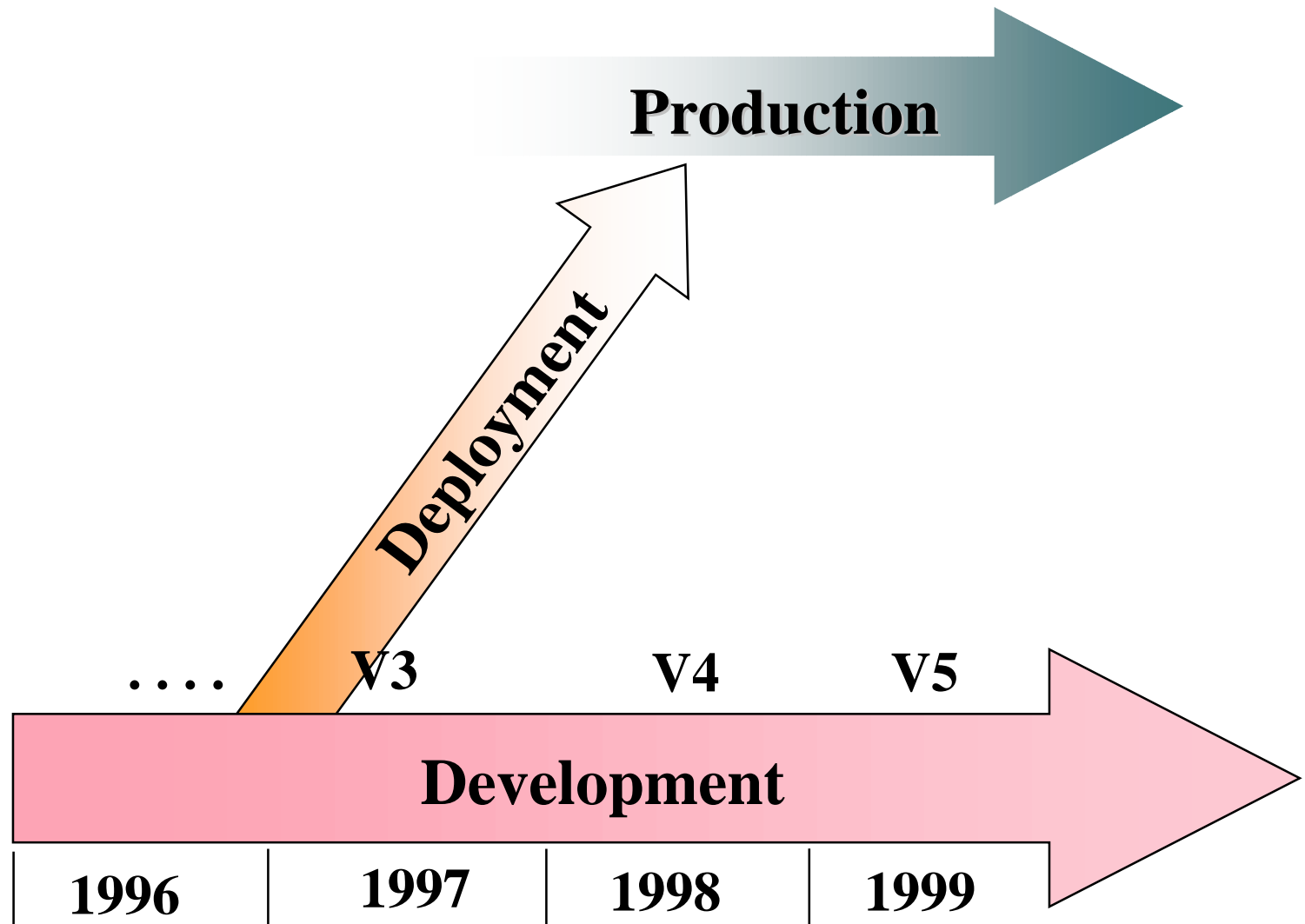
HPSS



Current Status at LANL

- Storage for Science-Based Stockpile Stewardship (SBSS) program, ASCI, grand challenge problems
- In production in both open and secure networks
- Currently accessible from Crays and ASCI Blue systems with deployment to LANs in progress

Current Status at LANL



Current Status at LANL

- Version 3.2 deployed, v4.1 Q4 1998
- Locally-written Parallel Storage Interface (PSI) is user interface
- Metrics

Users	Files	Storage	Growth
200	350K	20TB	33 GB/day

Storage growth is ~ 66 GB/day

Availability is ~ 95% since 1/1/98

Availability problems primarily due to other causes than HPSS: HIPPI network, DCE server

Current Status at LANL

- Data Transfer Performance
 - 500KB/s for reads and writes of small files to disk
 - 10MB/s for reads and writes of large files to disk
 - 10MB/s for reads and writes of large files to one-way tape
 - 20MB/s for reads and writes of large files to two-way tape
- Networks: FIDDI for control,
HIPPI-800 for data transfer

Current Status at LANL

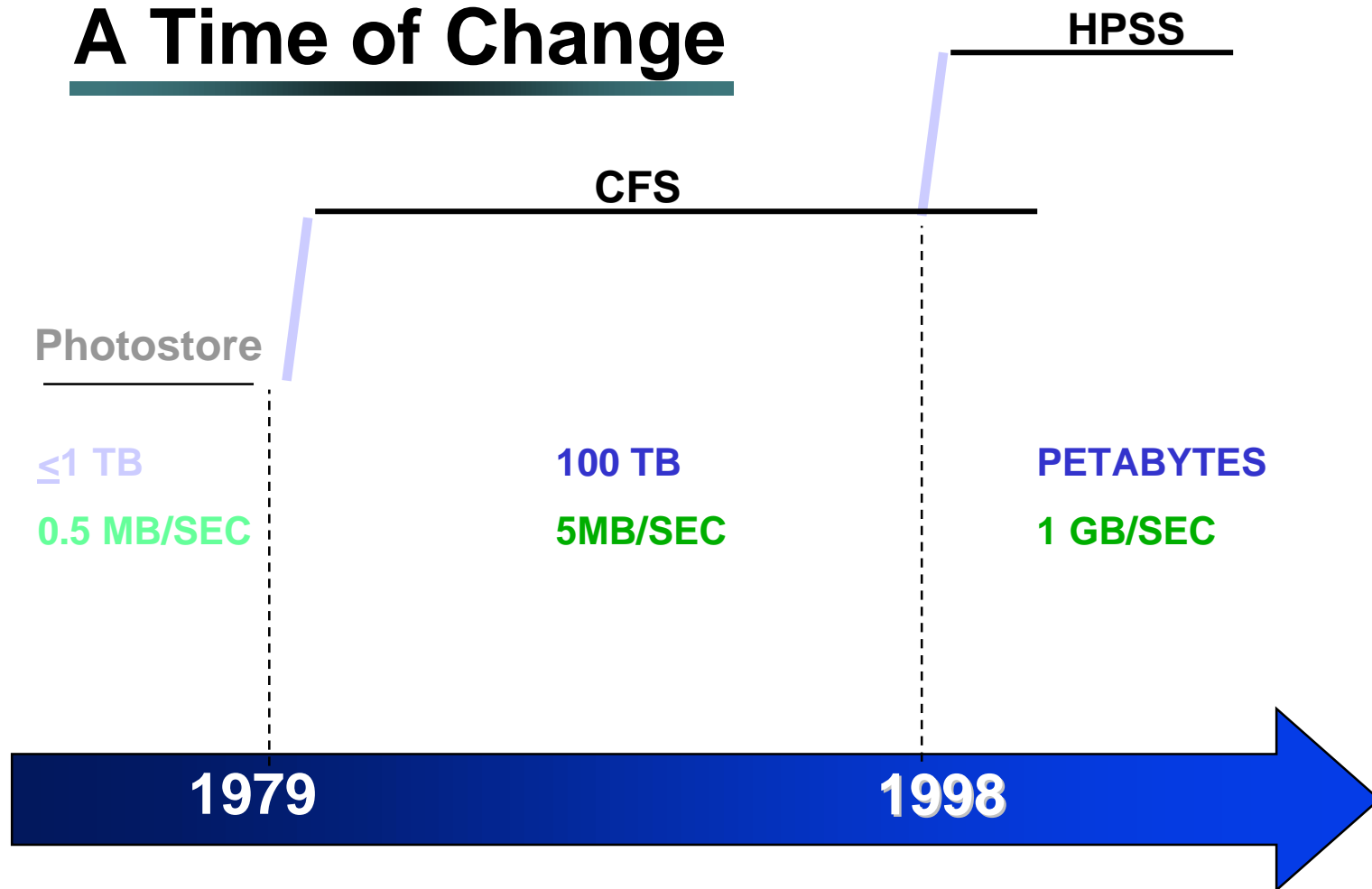
- Performance Issues
 - Small file performance: file and metadata creation - currently 1+ sec
 - Disk performance: problems with SSA adapter
 - System configuration: limited equipment funds
 - HIPPI device driver problems
- File Size Issue - much smaller than expected
 - open: 67MB, secure: 38MB

ASCI Requirements

- Accelerated Strategic Computing Initiative (ASCI)
 - Purpose - to accelerate computing technology
 - Funded by DOE
 - Replace nuclear testing with modeling and simulation

ASCI Requirements

A Time of Change

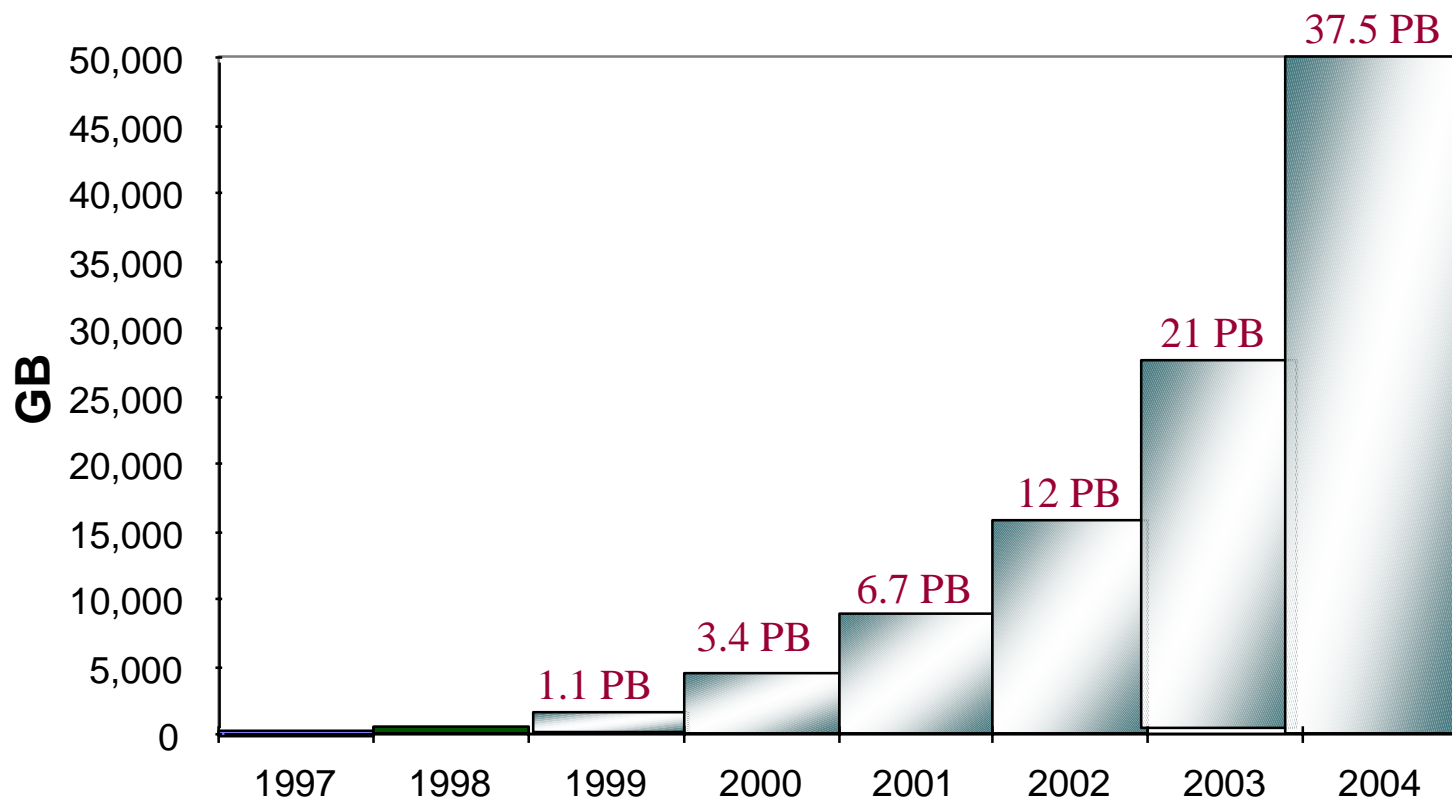


ASCI Requirements

- Two Driving Assumptions
 - Capacity: 750 Memories/year
 - Growth Rate
 - Bandwidth: 1/2 of memory in ≤ 20 minutes.

ASCI Requirements

The ASCI Data Storage Challenge (ASCI System Memory & Storage Growth)

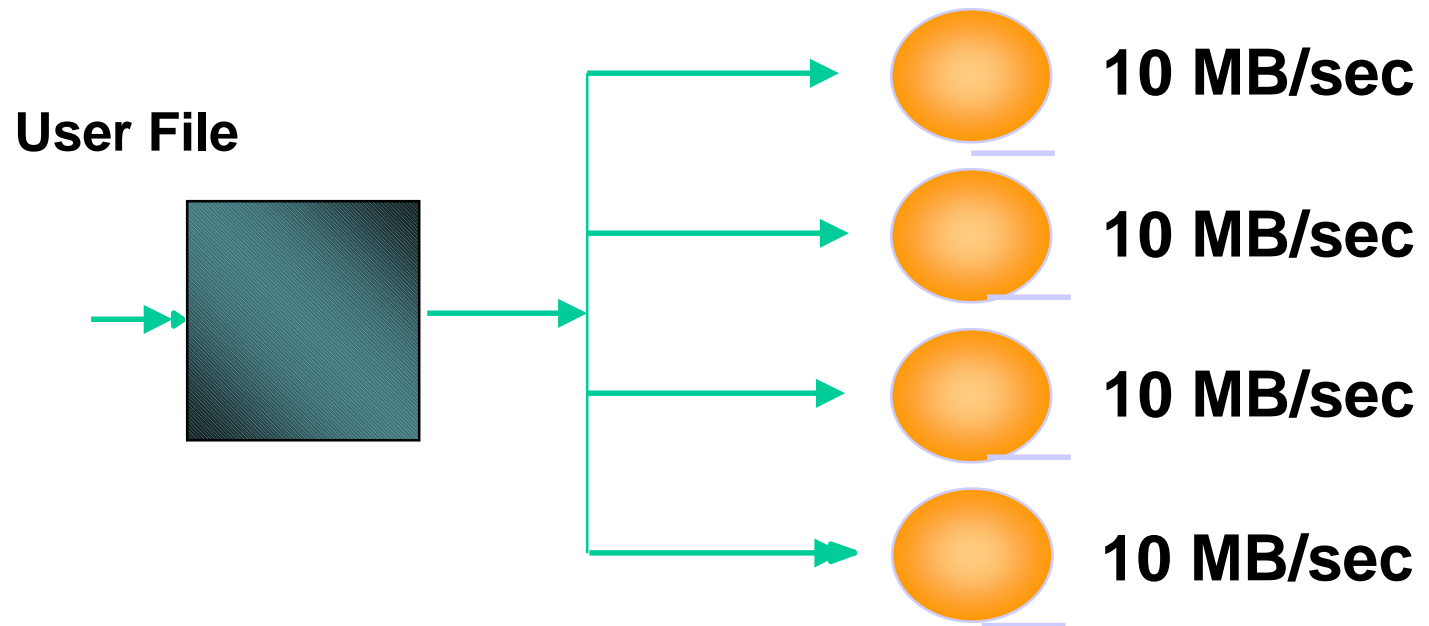


Challenges

- Funding for data storage to meet ASCI needs
- Accelerating data storage technology
 - network-attached storage devices (NASD)
 - striped tape systems (RAIT, RATS)
 - bandwidth aggregation devices (GNATS)
 - innovative caching, pre-fetch, data reduction techniques
 - practical, scalable, parallel I/O
 - practical, scalable, storage management
- New data storage paradigm

Challenges

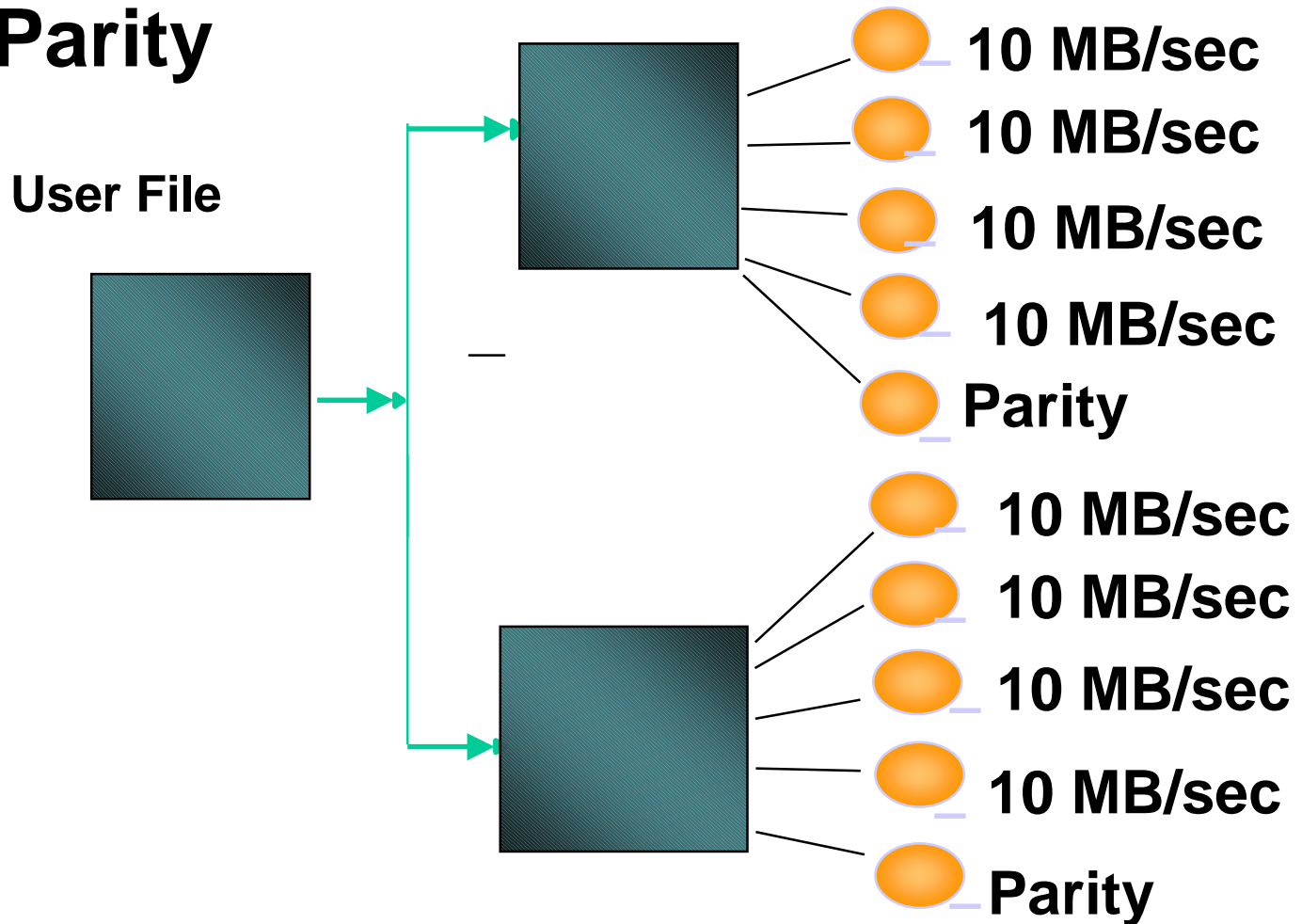
HPSS Tape Striping



Aggregate throughput: 40 MB/sec

Challenges

HPSS Multi-level Tape Striping with Parity

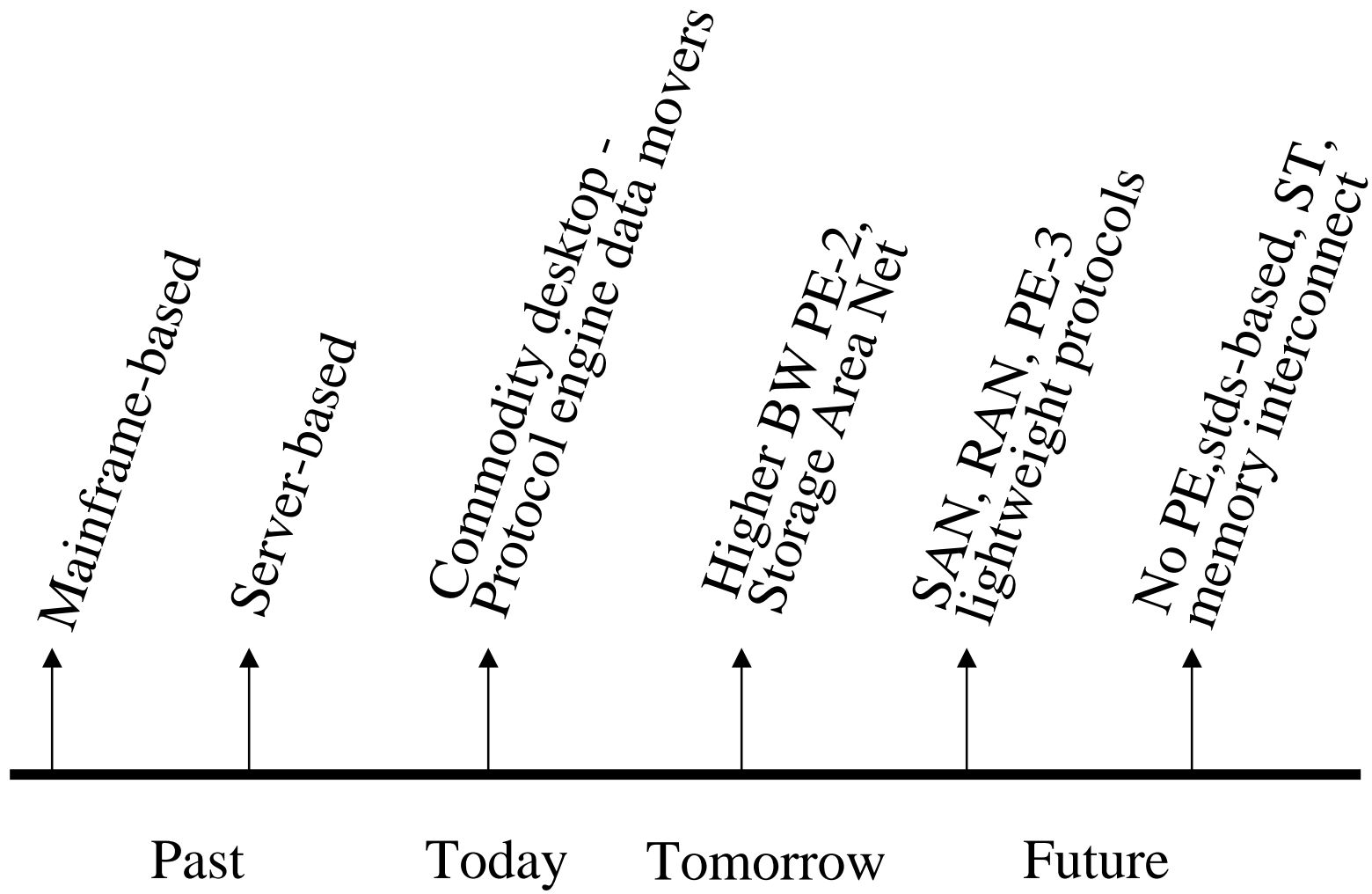


Vision

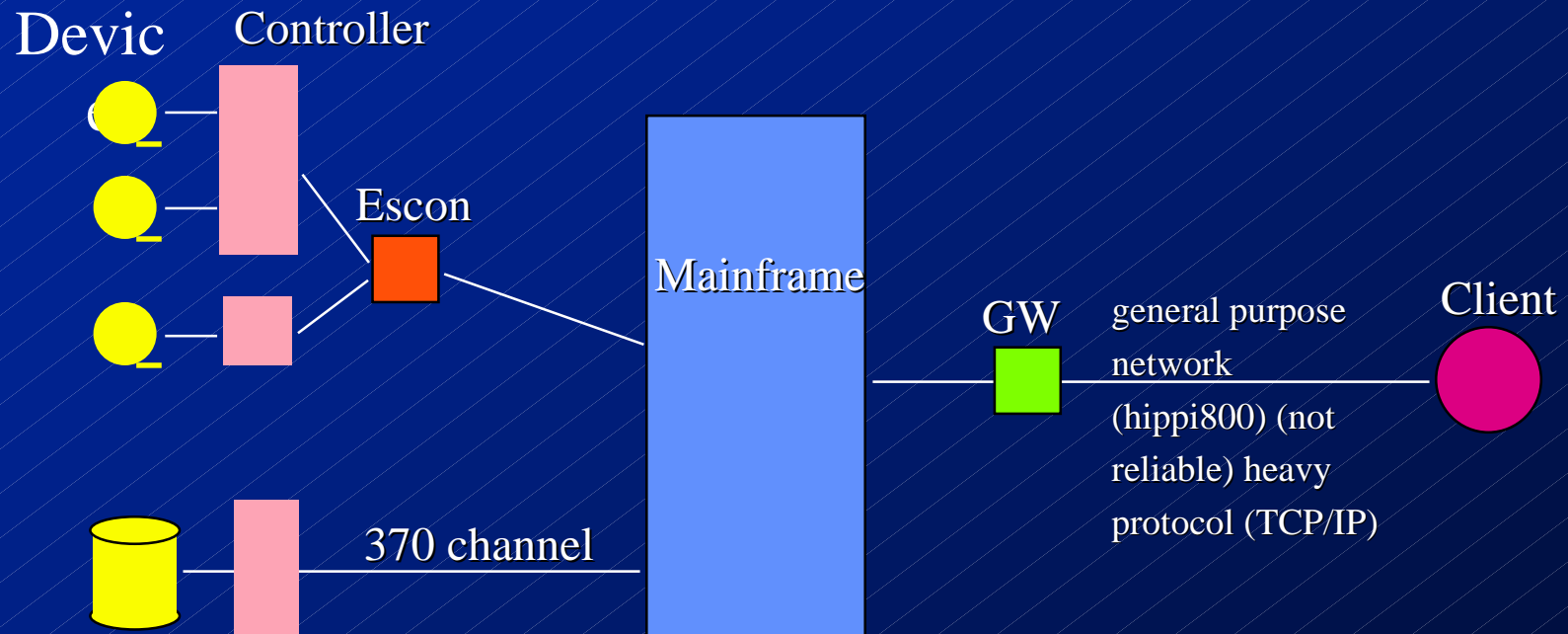
- Common Data Storage Infrastructure
 - Improved connectivity
 - Enhanced performance
 - Peripheral sharing
 - Central administration
 - Higher device utilization
 - Increased availability

Vision

Evolution of Data Storage Infrastructures

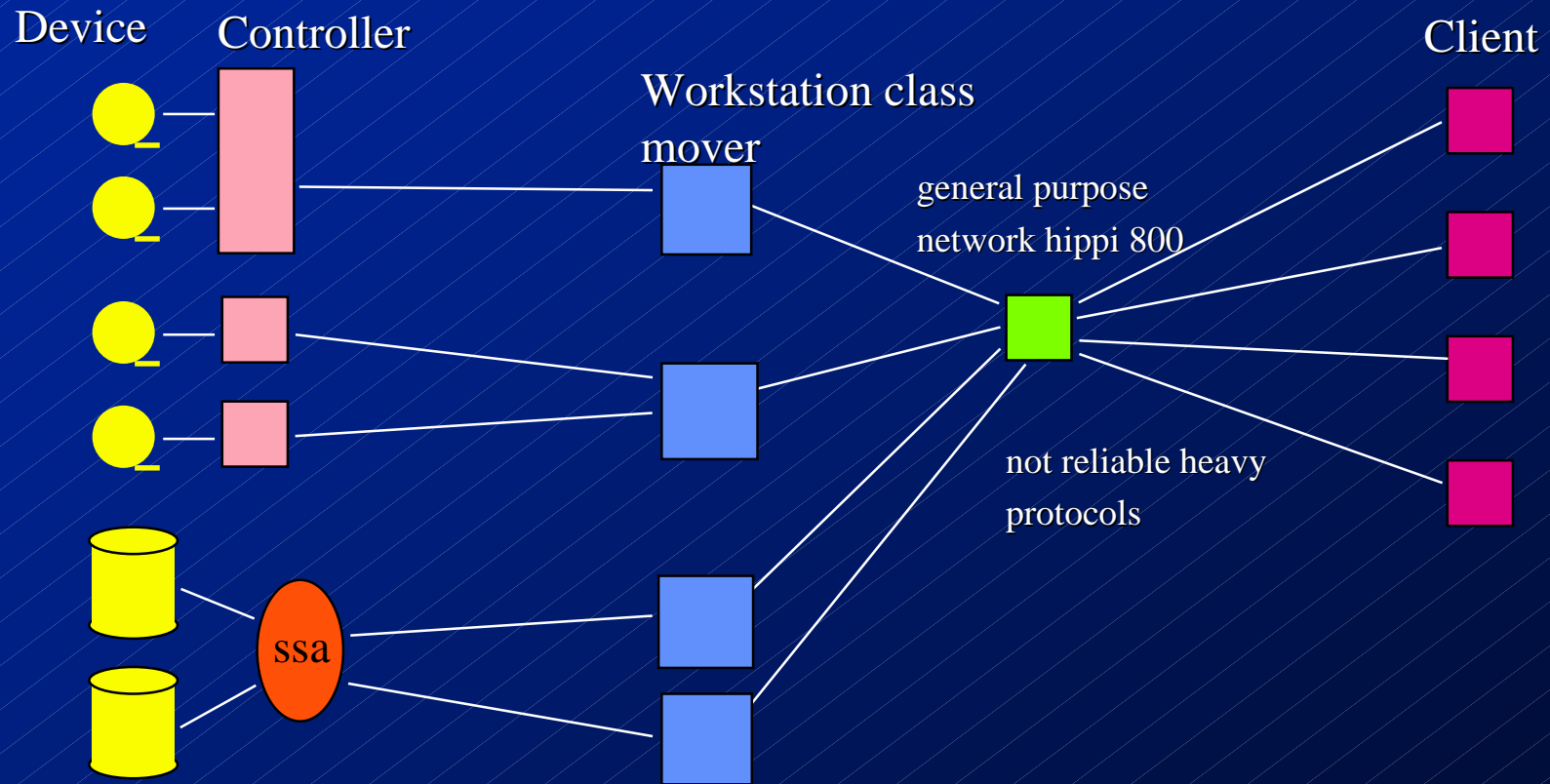


Past Mainframe-Based Movement



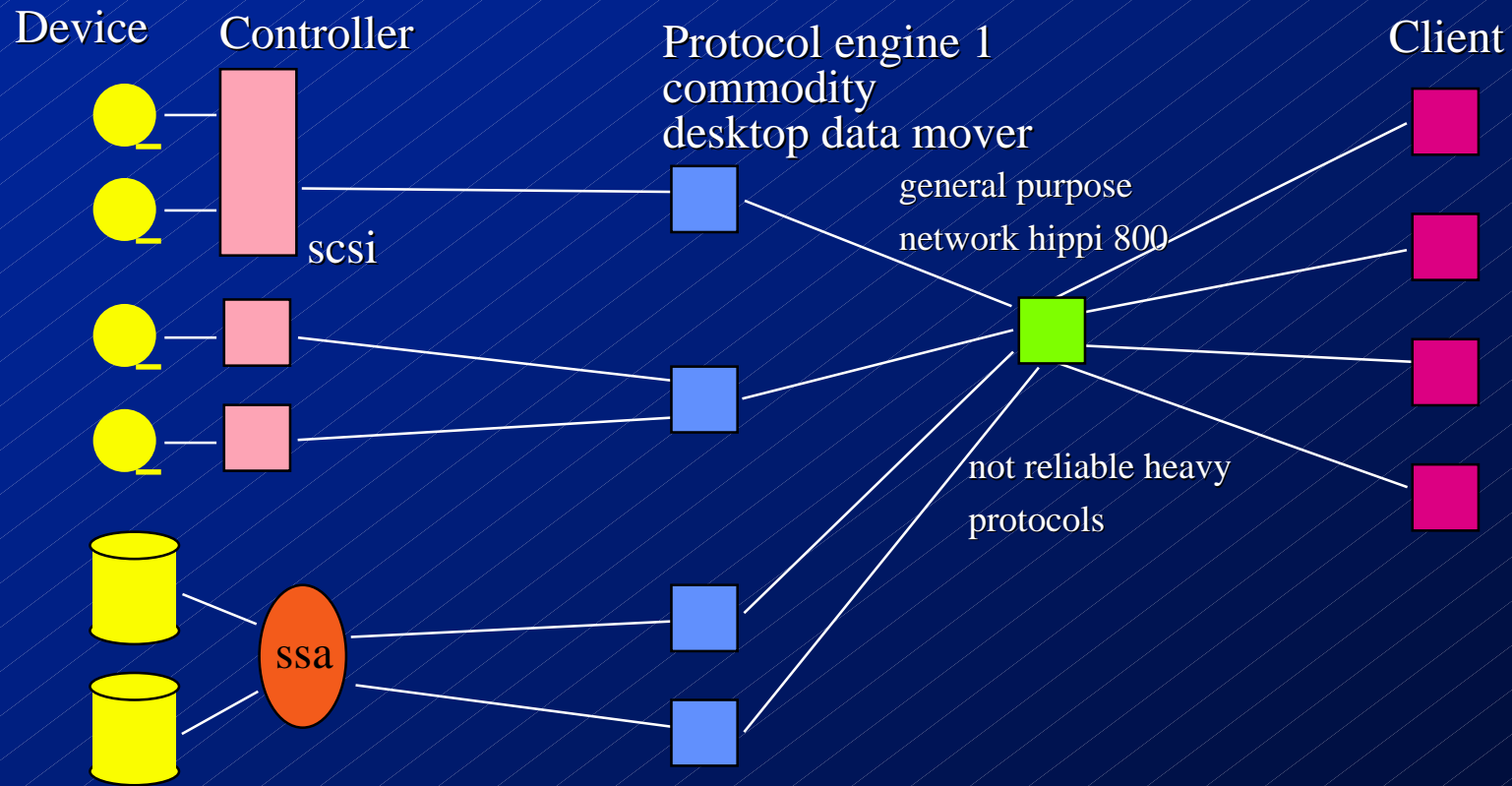
Data moves from client through general purpose net to mainframe, through 370/escon storage network to peripheral, storage network is just a set of channels from mainframe intermediary to peripherals, escon allows for switching and sharing between mainframes.

Recent, Past, Workstation/Server Class Data Movers



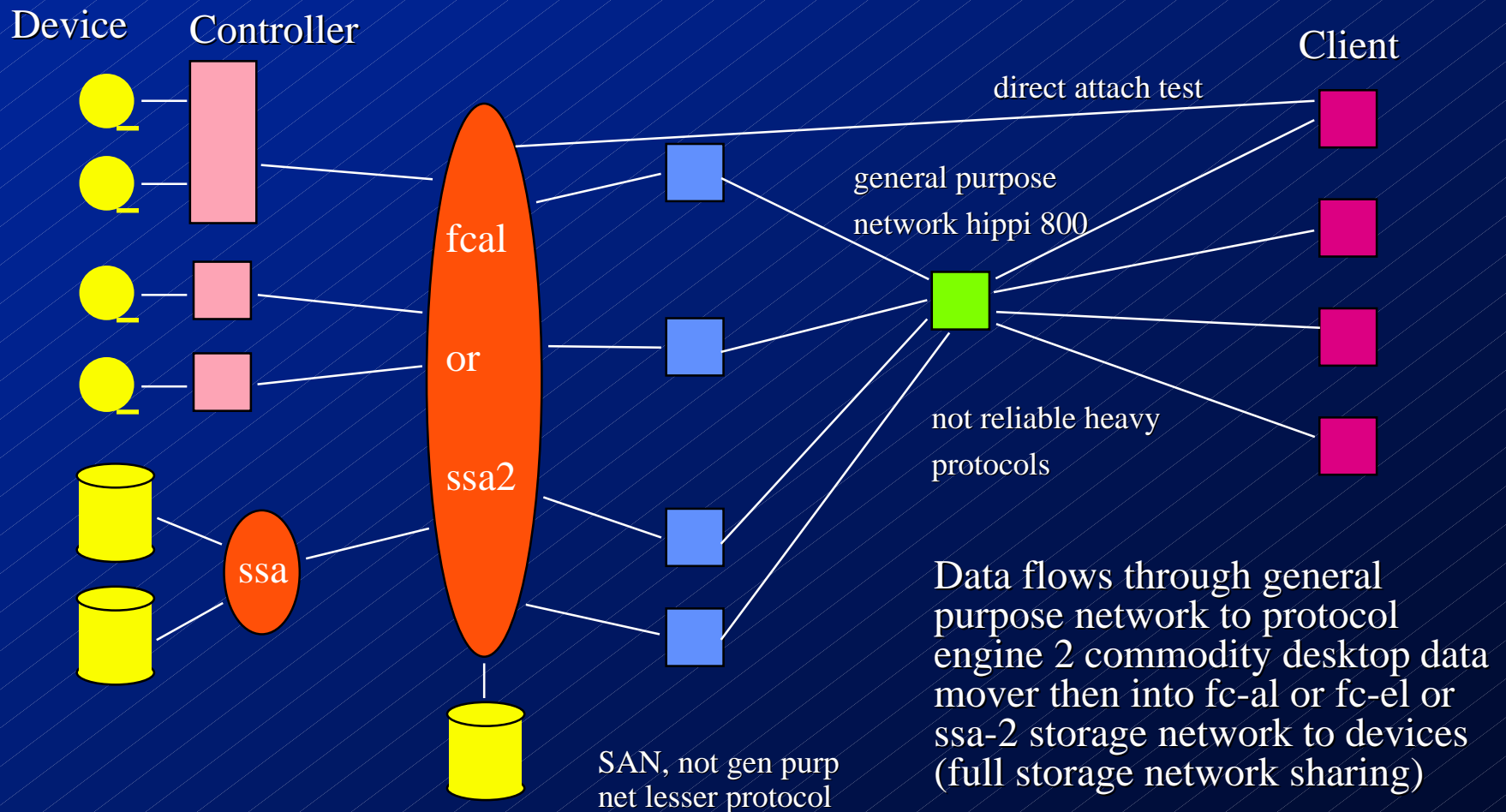
Data flows from client to mover via general purpose network then scsi/ssa to devices with or without sharing scsi - channel and ssa - storage network

Current - Commodity Desktop Protocol Engine Version 1 Data Mover

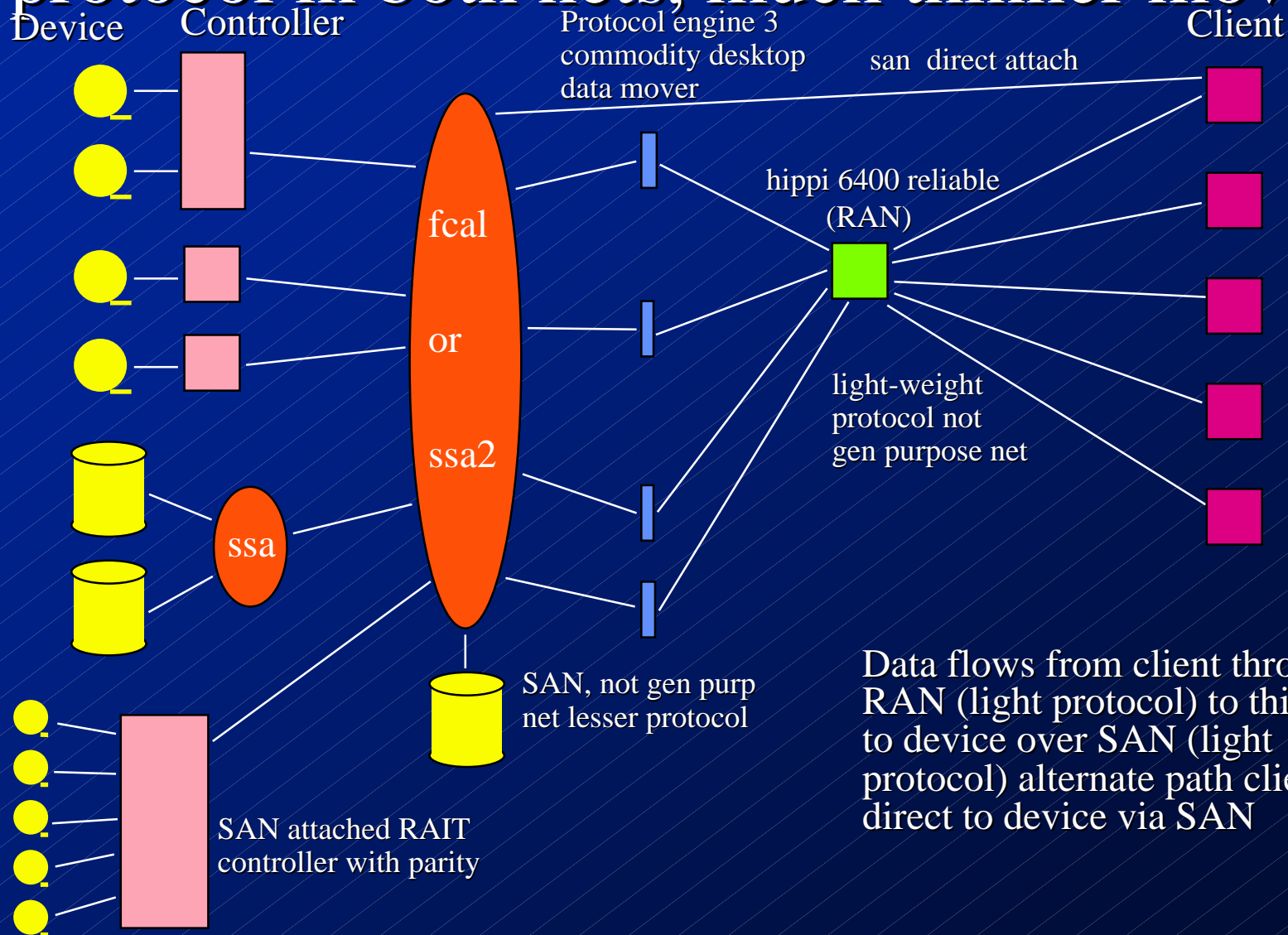


Data from client to mover via general purpose network then through scsi/ssa to devices with or without sharing scsi - channel and ssa - storage network

Next: Higher bandwidth mover to device with standards-based sharing, begin direct client attach experiments

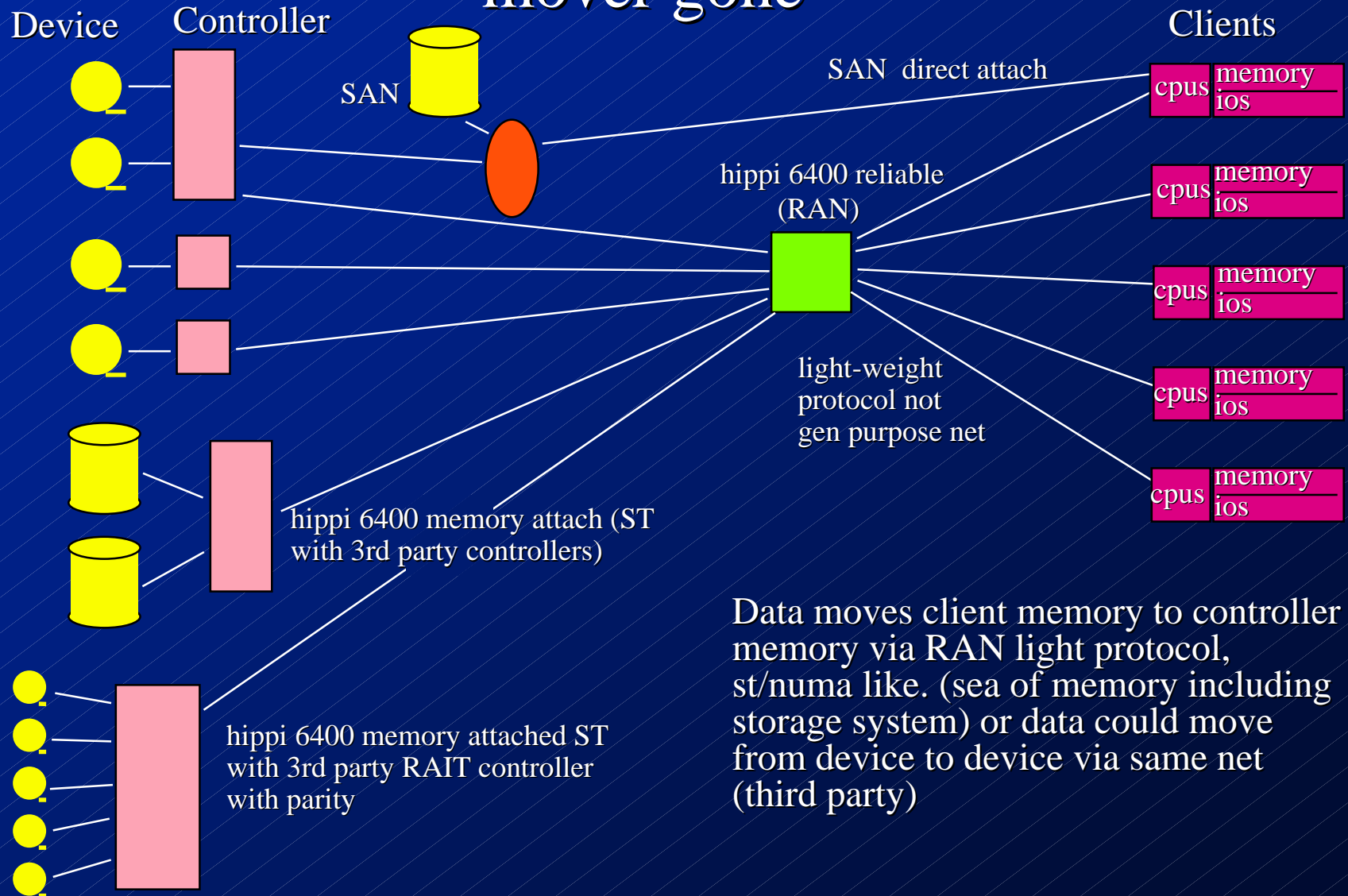


Beyond: SAN for mover device connectivity, RAN for client to mover, light weight protocol in both nets, much thinner mover



Data flows from client through RAN (light protocol) to thin mover to device over SAN (light protocol) alternate path client direct to device via SAN

Way beyond: no protocol engine, standards based controllers, ST with third party for device to device, memory interconnect instead of IOS interconnect, mover gone



Data moves client memory to controller memory via RAN light protocol, st/numa like. (sea of memory including storage system) or data could move from device to device via same net (third party)

For more information on HPSS

<http://storage.lanl.gov/cic11/hpss.html>

email: hpss_help@lanl.gov