APE Systems Parallel I/O Architecture

Piero Vicini
for the APE Collaboration

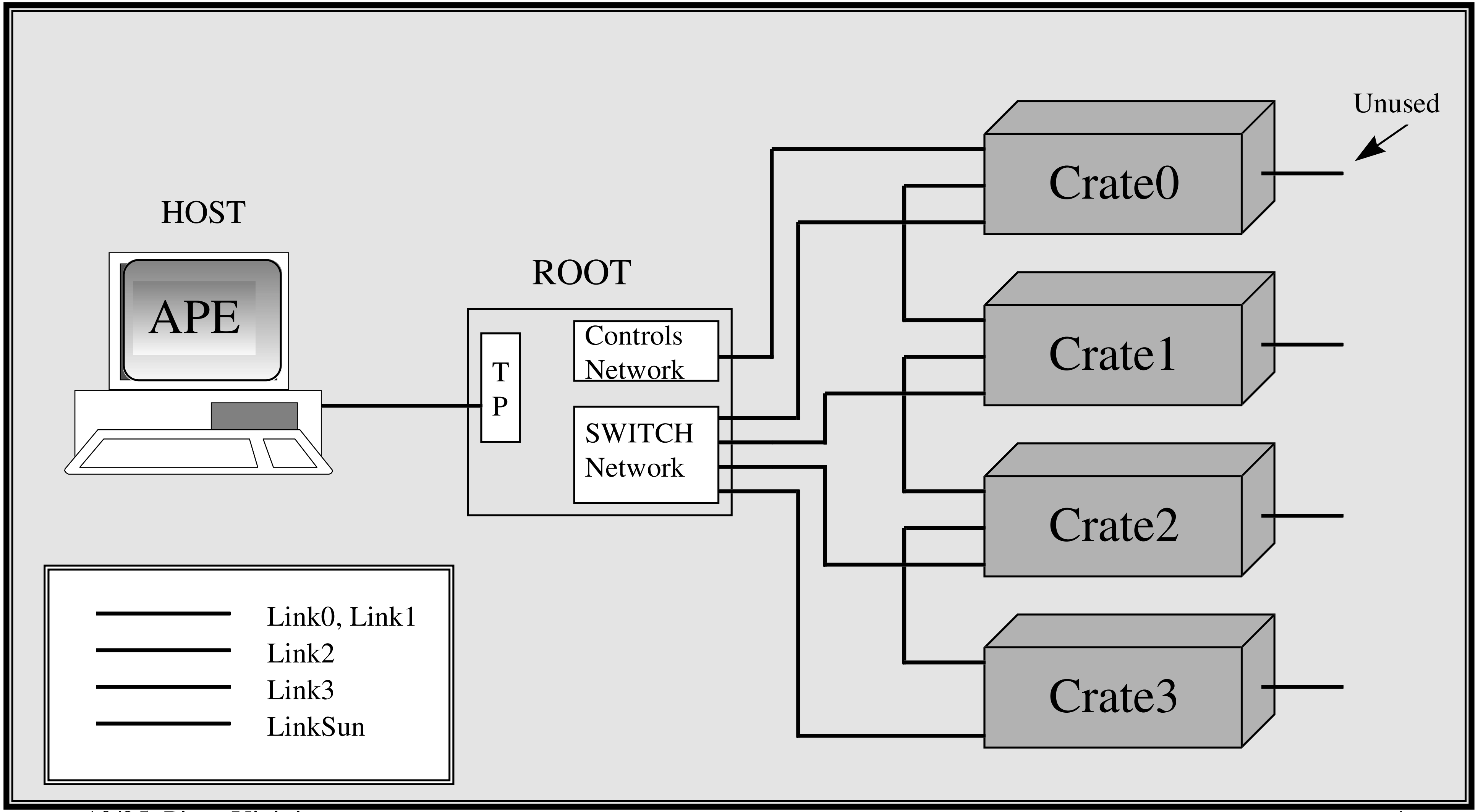
Main Topics

- APE100 Systems I/O Architecture
- Implementation of Fast I/O device using HIPPI protocol
- APEmille parallel I/O

APE100 I/O (a short review..)

- The *Host Interface* of APE100 systems is based on a Transputer network.
- Through this network is possible to download and upload data and program to/from any device of the whole machine.
- I/O performance ~ 300 kB/s

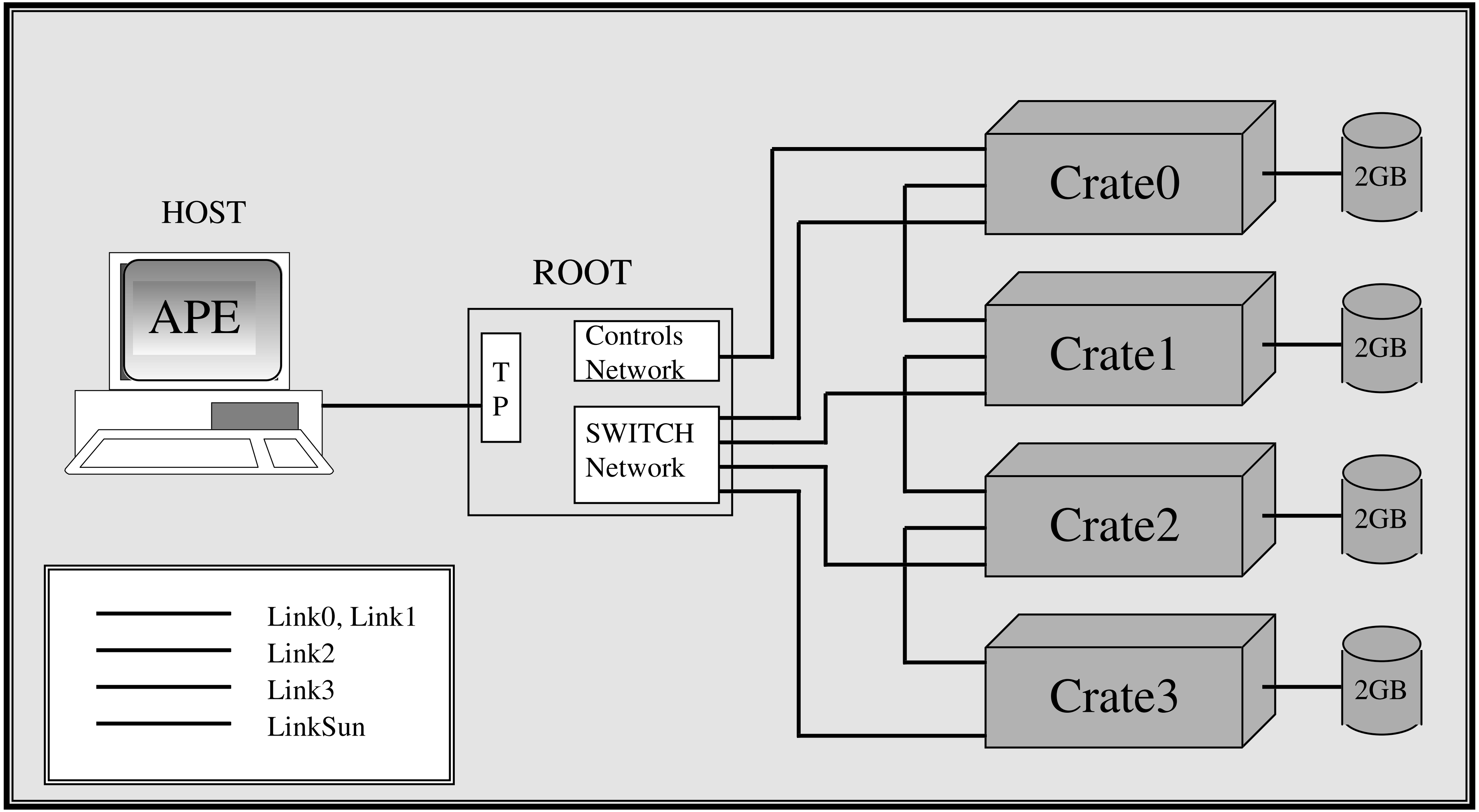
APE100 Host Interface



Local Disks

- In order to improve over all performances it was developed an add-on board based on a SCSI/Transputer Interface.
- Every Crate has its own disk directly connected to its Controller Transputer.
- Users could declare “Disk” variable treating them as a “Register” variable. Limitations are present due at extremely low speed of disk devices....

APE100 Local Disks



Local Disks

- Benefits:
  - Virtual memory.
- Limitations:
  - There is no file system on the local disk. The higher level software uses this kind of access to implement a data vault system.
  - Disks are not accessible from an external device. The I/O performances are fixed by the Host/Root/Crate network speed.

Performances

- Peak performances ~ 1.3 MB/s per crate
- The performance of such mass storage system is tightly depending from:
  - size of single burst transfer; i.e. number of words to transfer with a single system service;
  - time between successive transfers.

Performances (an example...)

- I/O transfer of 550 MB executed with 24 consecutive single burst transfers of ~ 23 MB each:
  - Performance ~ 1.0 MB/s per crate (= 77% of peak)
- I/O transfer of 180 MB executed with 540 consecutive single burst transfer of size ~ 0.33 MB each:
  - Performance ~ 0.2 MB/s per crate (= 16% of peak)

Why Fast I/O?

- Users require more I/O capabilities...
- Fast I/O implemented on standard protocols means “Open Architecture”.

FP Operations versus I/O considerations

- Estimated ratio (R) of FP operations versus I/O bandwidth (in words unit) in the hypothesis $T_{I/O} \sim T_{Calculus}$:
  - QCD $R < 70000$
  - Fluid Dynamics $500 < R < 2000$
  - Image Processing $200 < R < 50$
  - Virtual memory $100 < R < 25$
- QCD applications:
  - A typical (now running) QCD application requires today (local disk) 12 hours of calculus and 8 hours of disk operations. Extended lattice application should require 15 hours of calculus and 25 hours of disk I/O !

APE100 Fast I/O Hardware requirements

- Standard protocols
- High bandwidth
- Simple protocols -> High efficiency
- Switching technology
- Chip set available
- Off the shelf technology

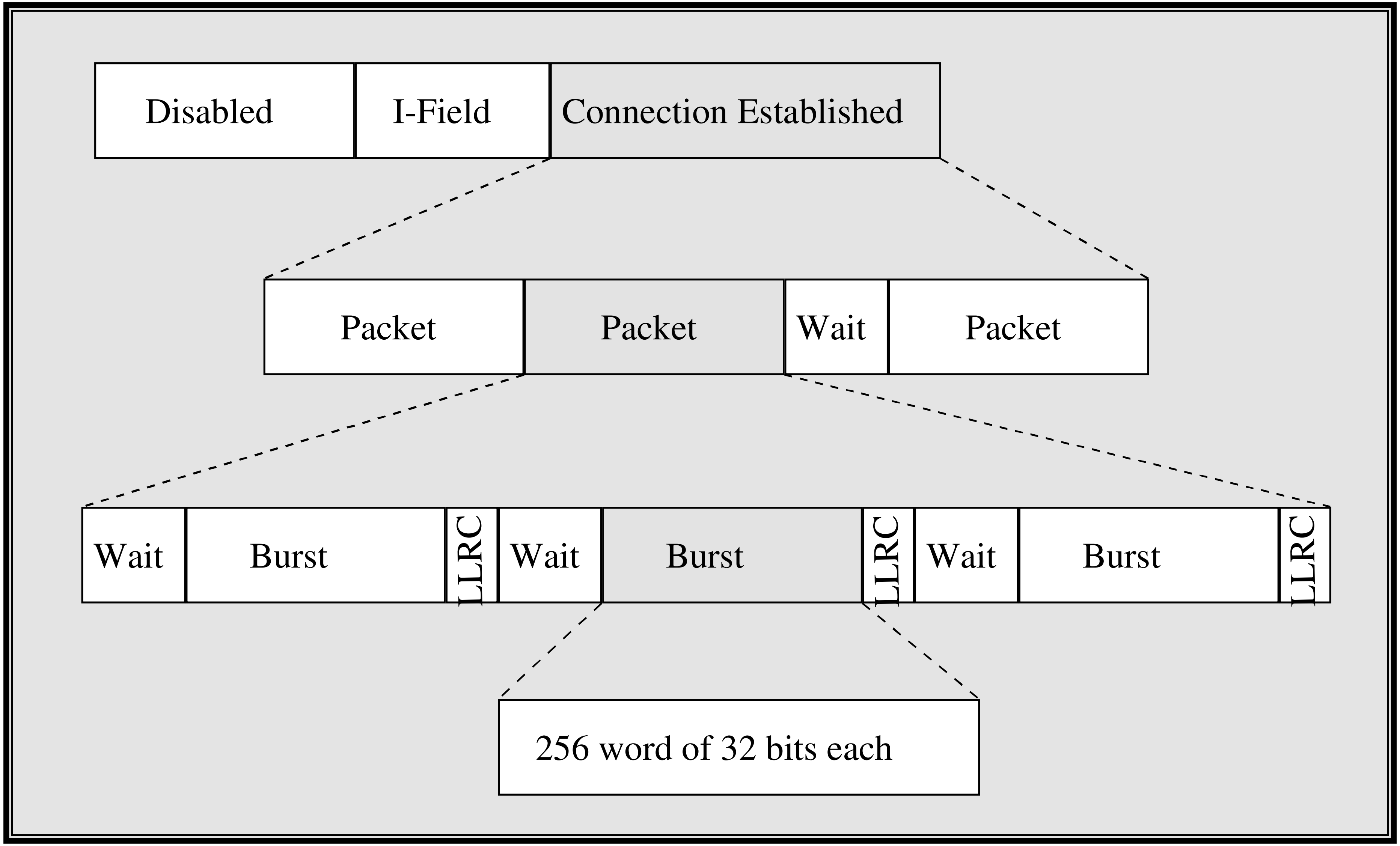
Selecting the right technology...

- Transputer Network upgrade (T800 -> T9000).
  - Not enough fast, not yet available...
- ATM standard:
  - network oriented (payload concept)
  - today performances 155 Mbit/s, up to 10 Gbit/s (2010!)
- FCS standard:
  - channel oriented
  - optical links required, performances up to 1.0 Gbit/s
- HIPPI standard:
  - channel oriented supercomputer standard
  - mature technology ("plug and play")
  - 800 Mbit/s e 1.6 Gbit/s, in the future Serial HIPPI

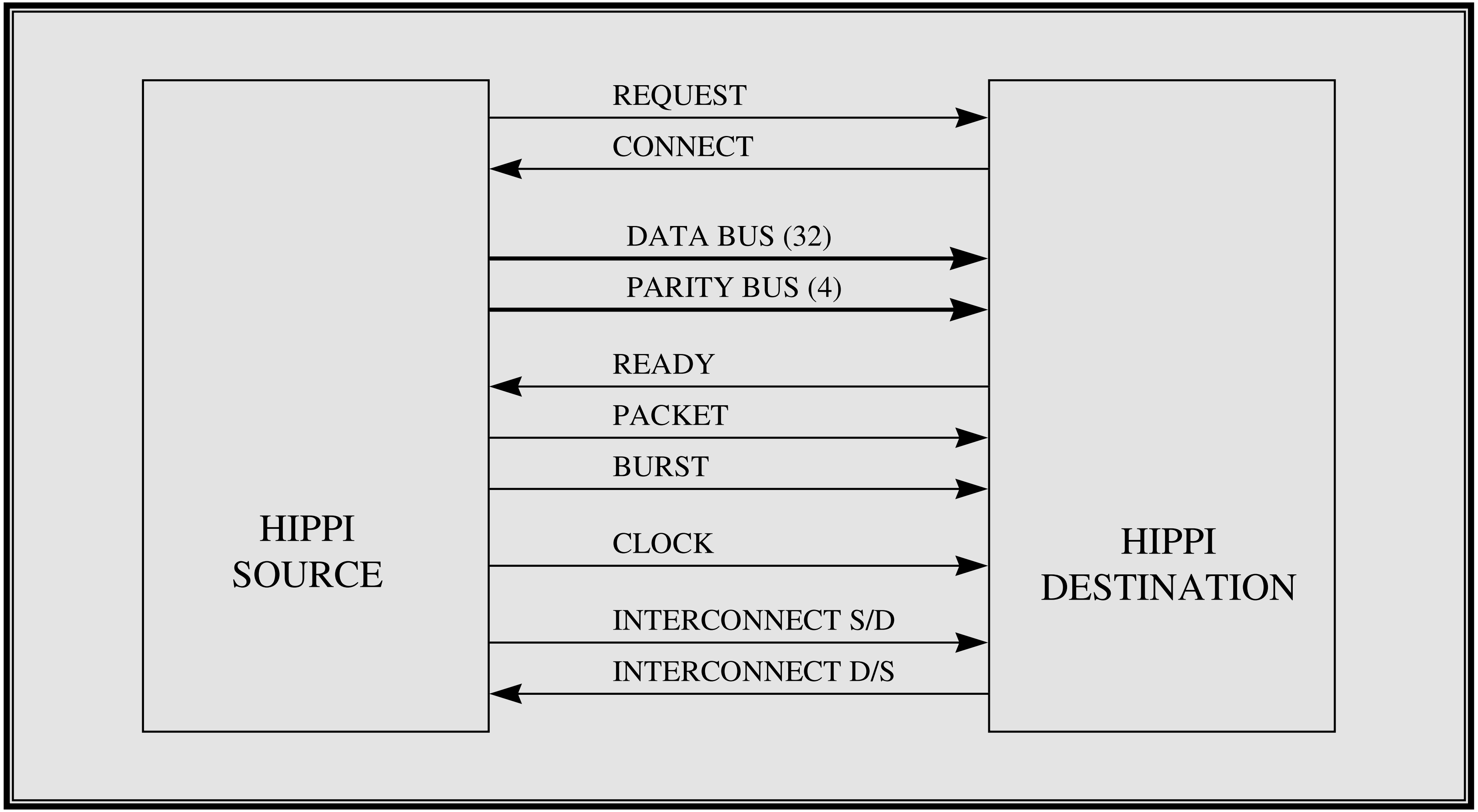
HIPPI

- HIPPI standard defines the physical layer of an efficient simplex high performance point to point interface operating at speeds of 800 or 1600 Mbit/s.
- Characteristics:
  - Point to point connections on copper cables up to 25 m. Two HIPPI channels realize a full-duplex channel.
  - Data transfers are performed in increment of burst of 256 words each.
  - Simple signaling and control sequences (average transfer rates approach peak transfer rates).
  - Easy and efficient to use in a circuit-switched environment.

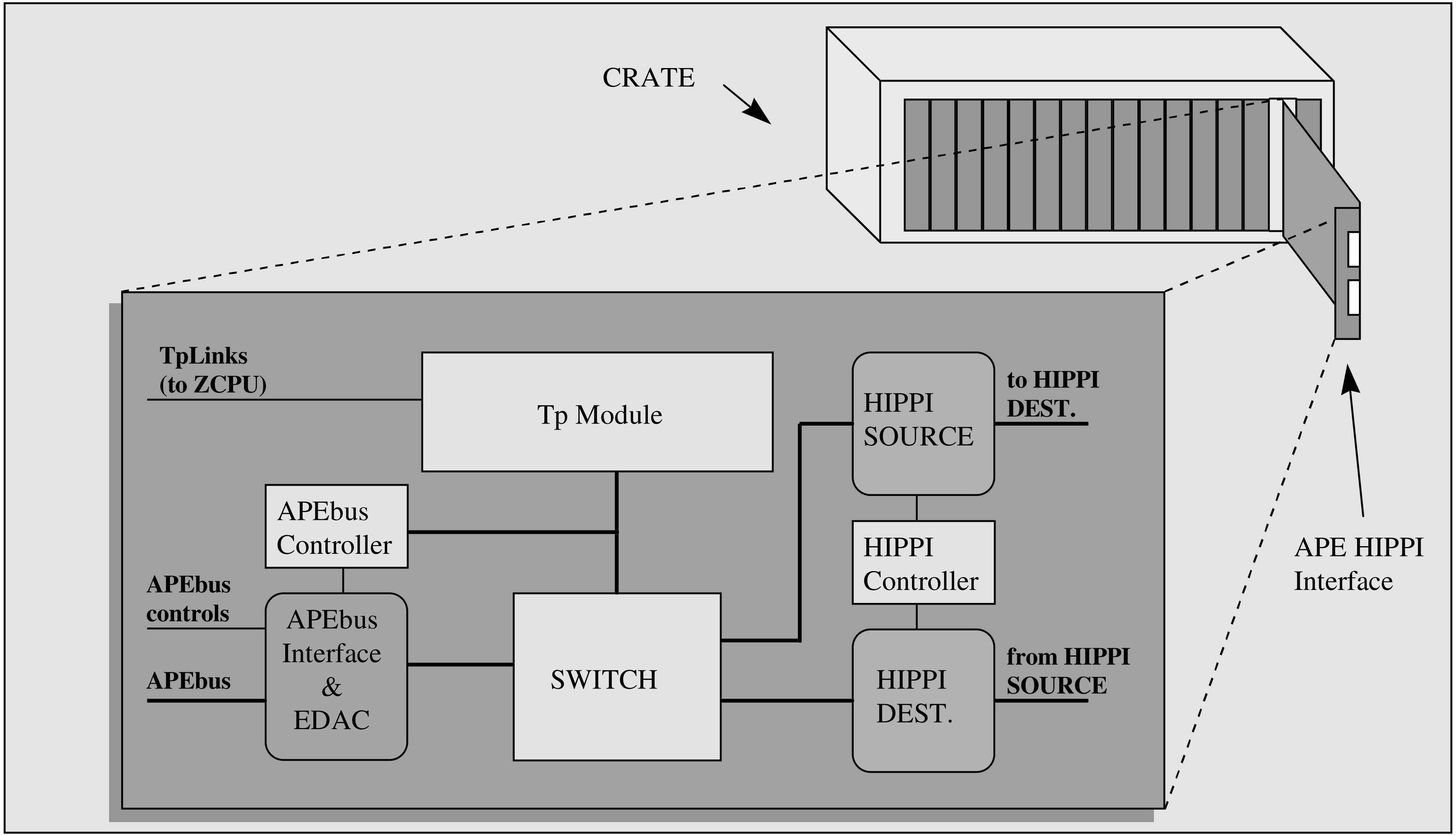
HIPPI Format



HIPPI Signals



APE HIPPI Interface

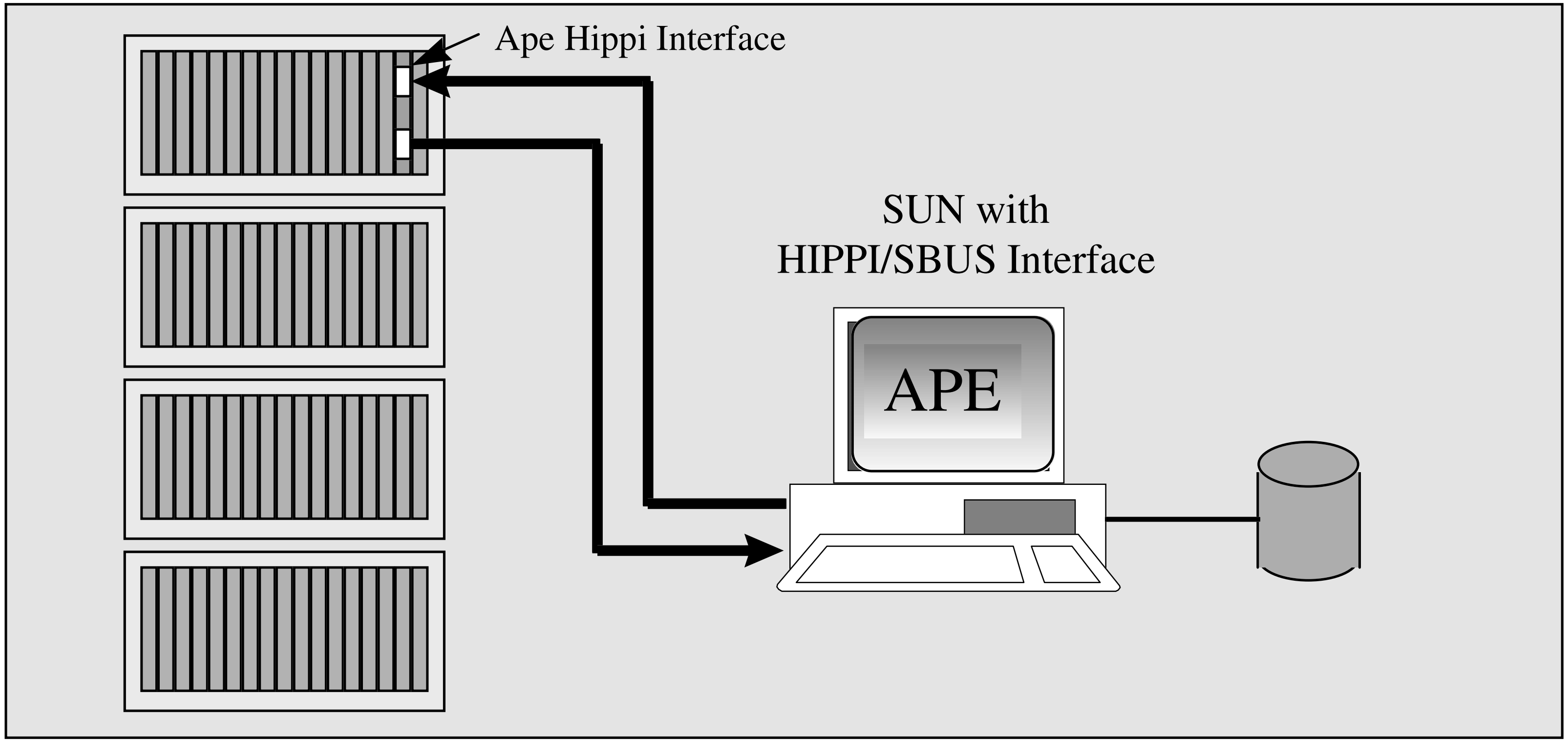


APE HIPPI Interface Mode Programmability

- In order to minimize transfer time, different memory access modes, executed in a single system service, are provided.
  - SINGLE NODE access:
    - single burst of memory access (of length L) from S (Start Address)
  - BROADCAST access:
    - L consecutive accesses to all nodes on a single board at memory location S
    - L consecutive accesses to the same node on all boards at memory location S
  - FORMATTED OUTPUT access:
    - Access to multidimensional data structures distributed on a slice of the nodes lattice. Output is equivalent to a sequential file.

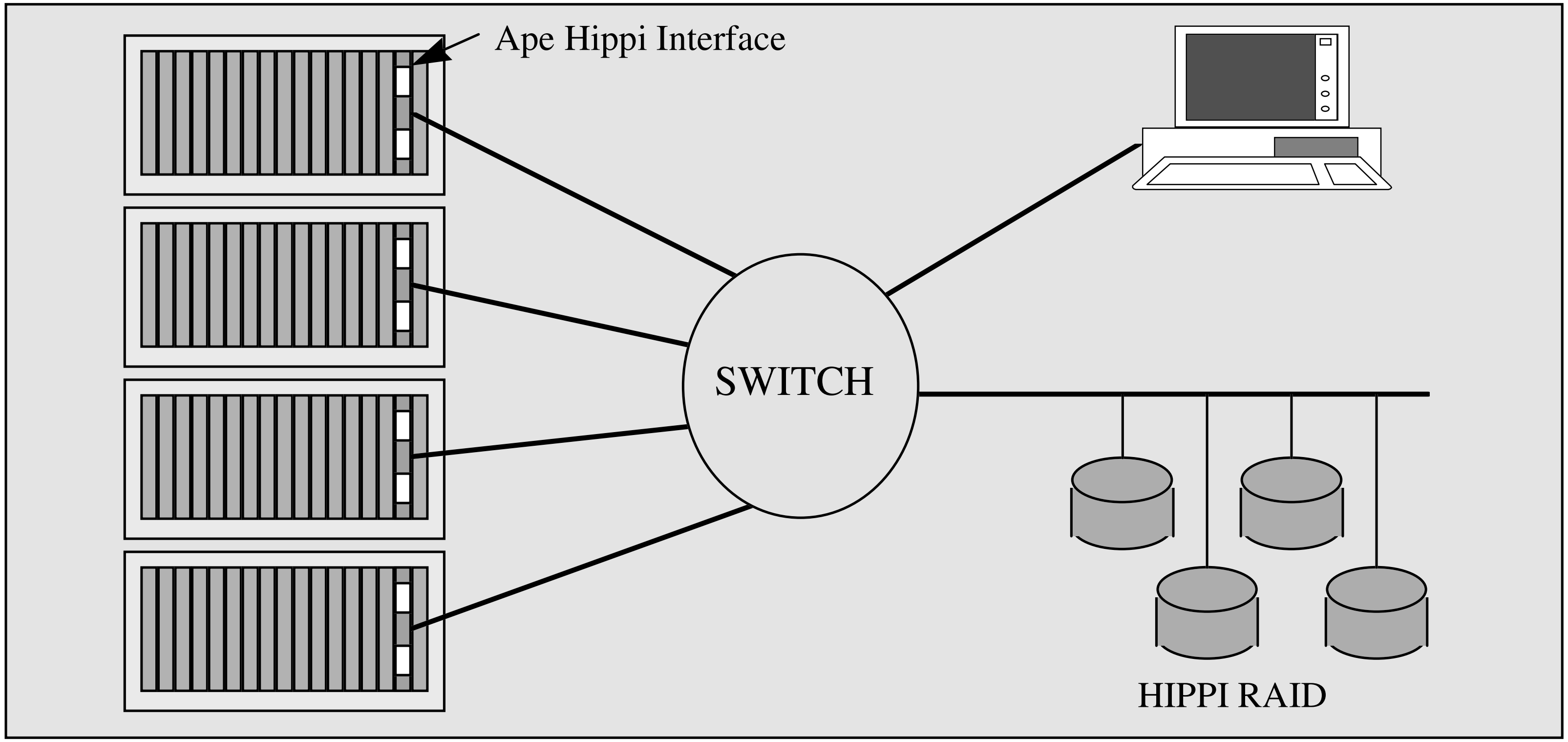
APE HIPPI System Architecture

- Option 1:
  - One APE HIPPI Interface per tower connected to an host server via HIPPI protocol. Data from other crates are provided via remote communications.

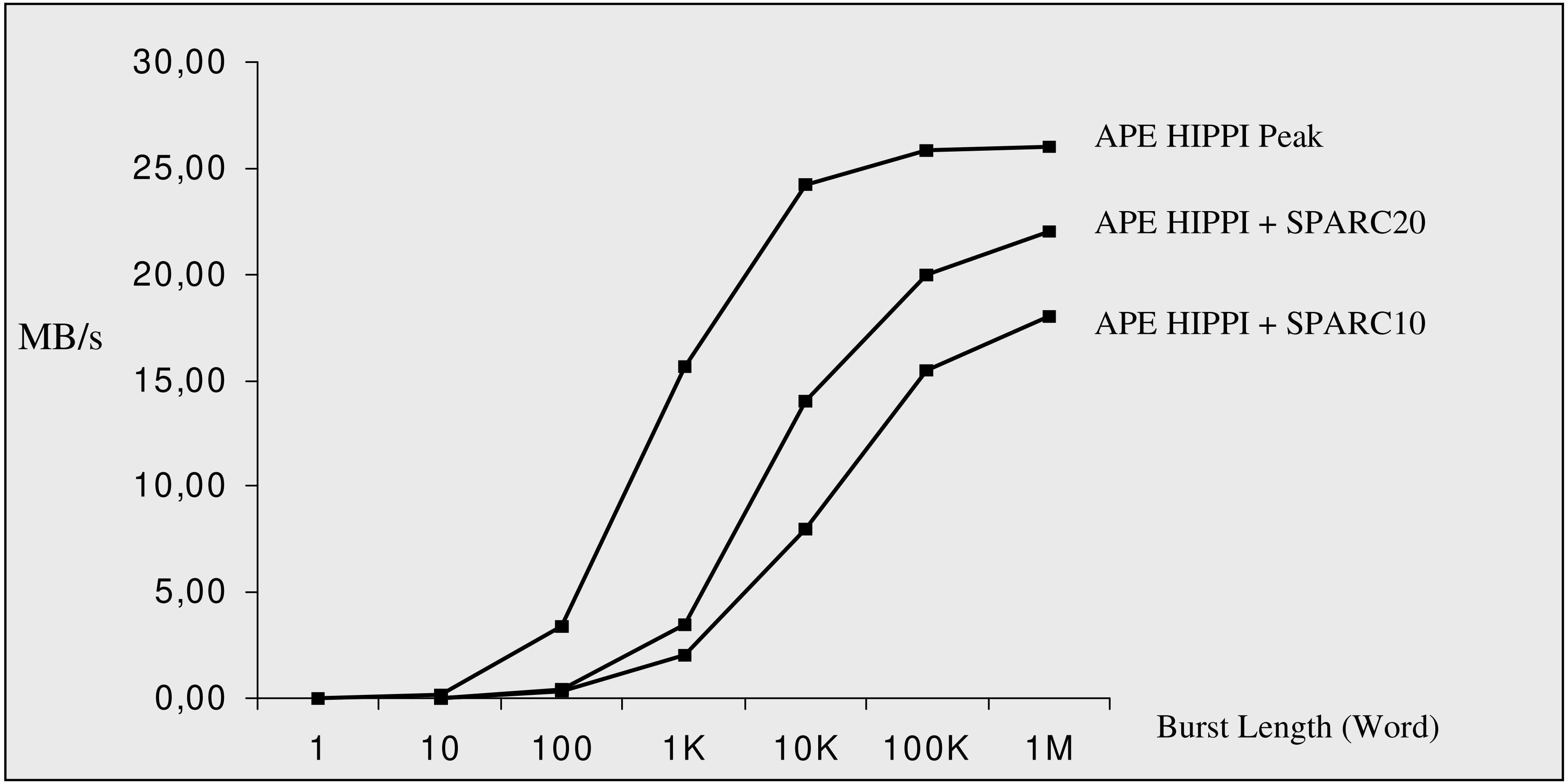


APE HIPPI System Architecture

- Option 2:
  - One APE HIPPI Interface per crate connected to an HIPPI disk server via an HIPPI SWITCH.



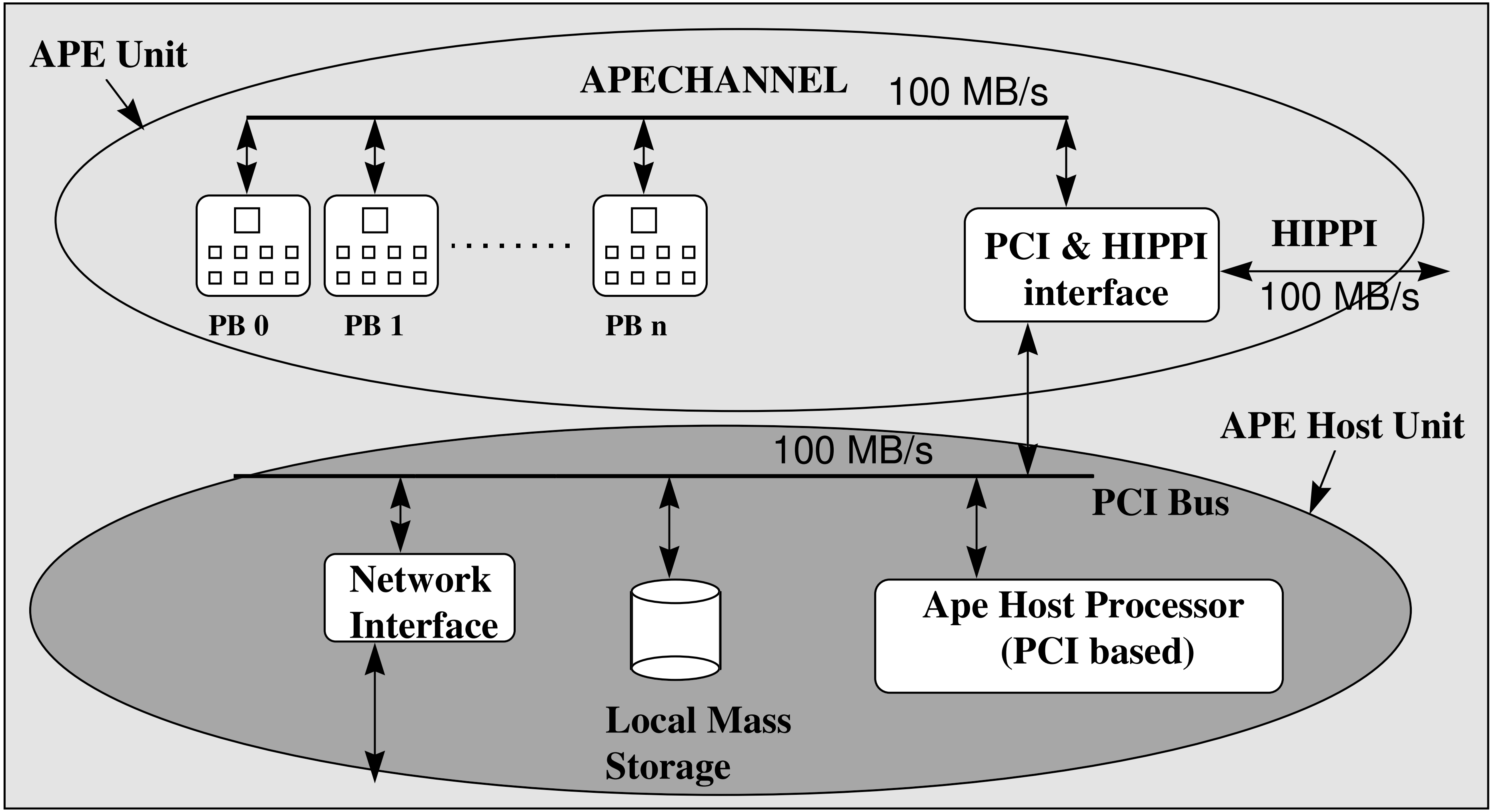
APE HIPPI System Performances



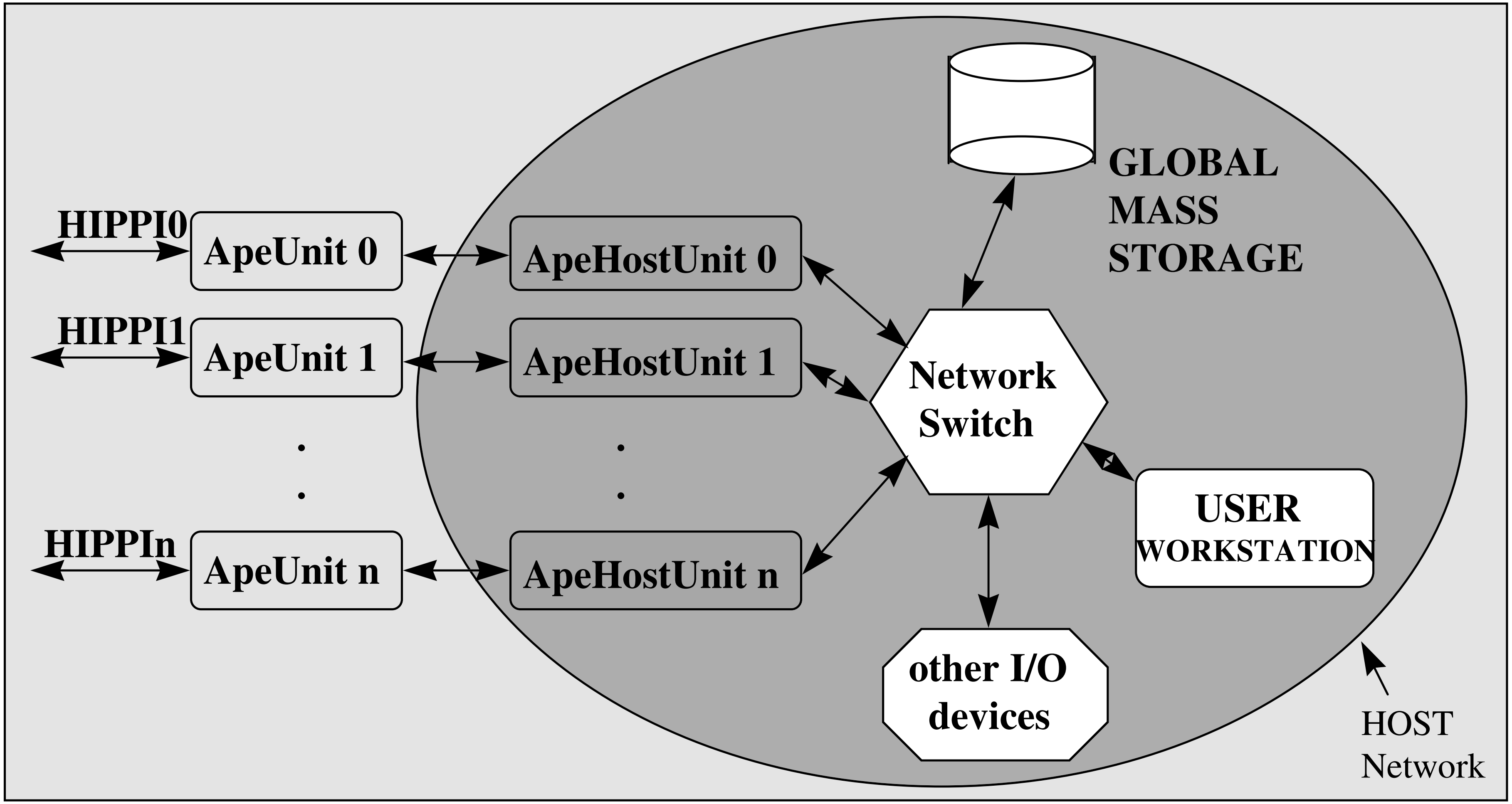
APEmille Host Interface

- APEmille Processing Board contains 8 FP nodes.
- A cluster of PB (up to 16 PB) is connected to (and controlled by) an APECHANNEL.
- APECHANNEL:
  - Bidirectional channel, 100MB/s (32 bit word length, 25 MHz).
  - Custom protocol allows burst transfers executed at full speed.
  - Technology independent architecture allows connections to standard bus (PCI) and to standard I/O channel (HIPPI)

APEmille Host Interface



APEmille Host Network



APEmille Parallel Fast I/O