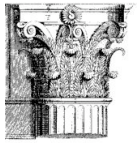
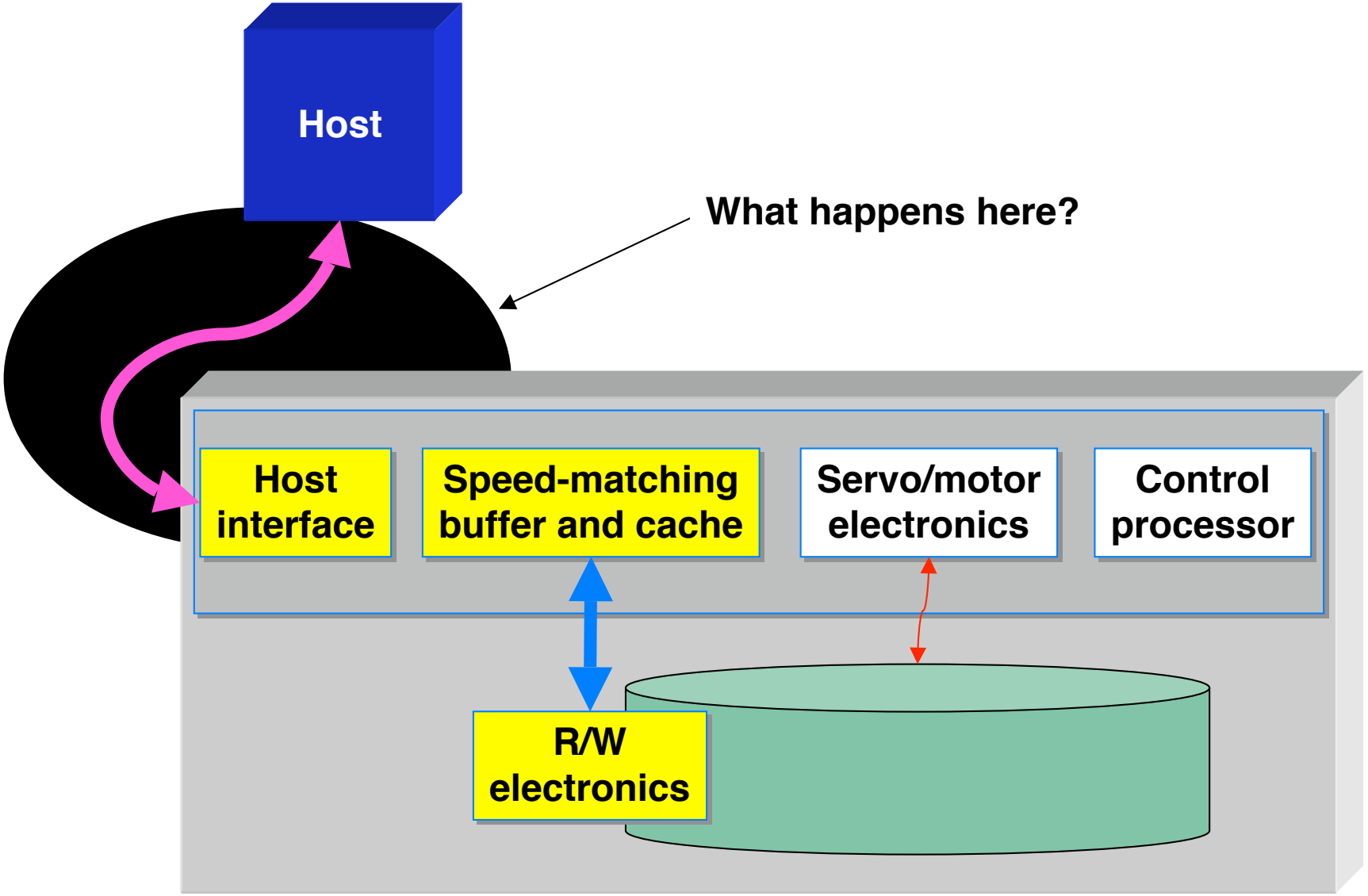
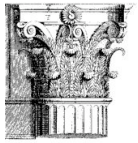
Storage area networks

- **Physical connectivity**
  - **links and loops**
- Command sets and protocols
  - what to say
  - how to send it
- LAN vs SAN
  - and other false dichotomies
  - case study: CMU NASD



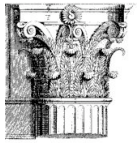
Storage-device interfaces





Storage interfaces: 3 levels

- **Physical connectivity**
  - what kind of bus/interconnect fabric
- **Command set**
  - how are requests formatted?
  - what requests can be sent?
- **Protocol**
  - buffer management
  - flow control
  - error recovery
  - security

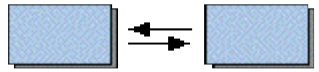


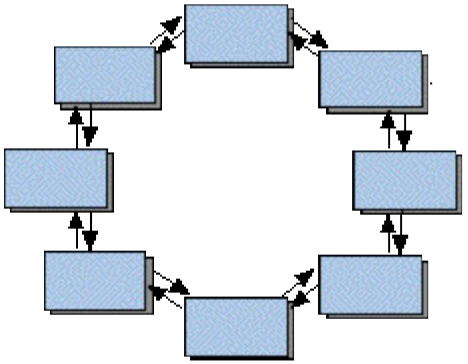
Storage interfaces: physical connectivity

- **Local attach**
  - **IBM channel (System/360 mainframe)**
  - **[E]IDE**
  - **parallel SCSI**
- **Shared attach -- SANs (Storage Area Networks)**
  - **parallel SCSI** (multiple initiators)
  - **IBM SSA**
  - **Fibre Channel**
- **Futures?**
  - **Ethernet**
  - **Infiniband**

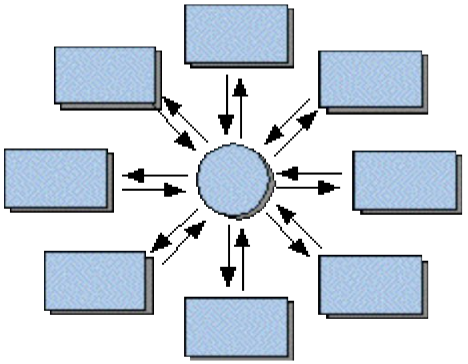
SAN topologies - FibreChannel

- **Physically:**
  - **100MB/s (800Mb/s)**
  - **point-to-point, loop, switched**
  - **copper: coax, or backplane traces**
  - **fibre: up to 500m multimode; 10km single-mode**

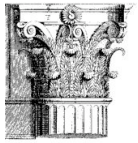




- **Components:**
  - **point-to-point links**
  - **hubs** (“wiring closet in a box”)
  - **switches**
  - **host- and device-adapters** (e.g., HP’s Tachyon chip set)

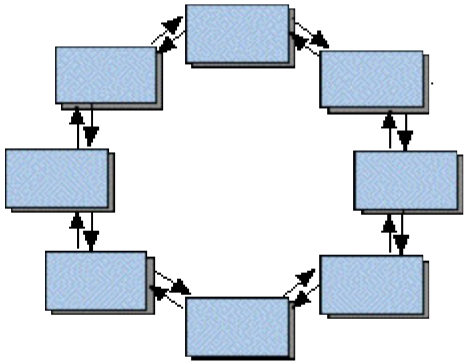


- **SCSI, IP encapsulations**



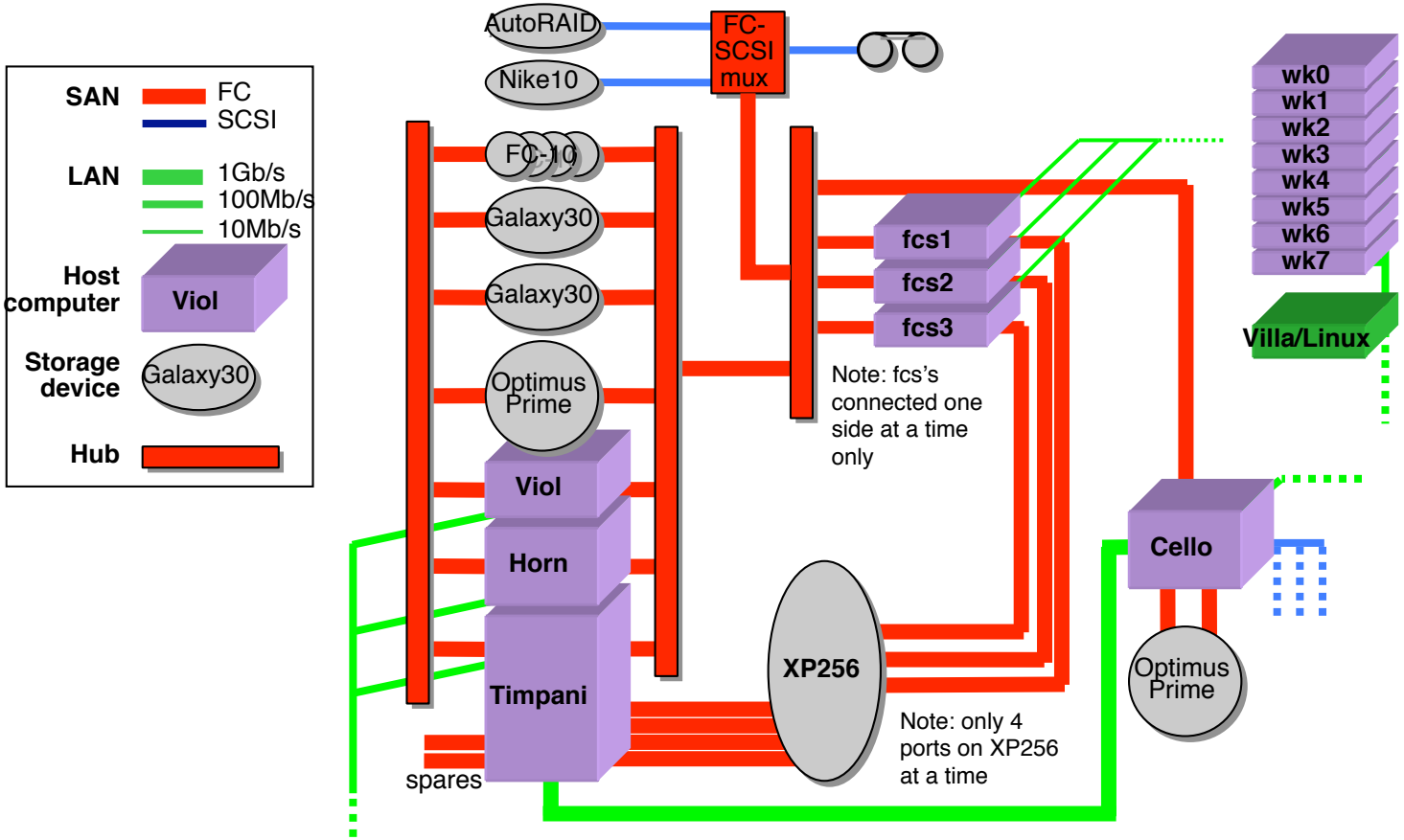
SAN topologies - FibreChannel arbitrated loop

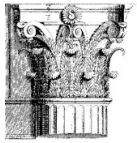
- A ring-connected structure
  - dual counter-rotating rings (failure tolerance)
  - hubs provide physical star-like topology
- Token-ring-like protocol: only one sender at a time
  - traversals:
    - lay claim
    - grant claim
    - transfer data
    - release



- Why bother?
  - low cost: a few $$ more than parallel SCSI
  - multidrop
  - supported now by disk drives

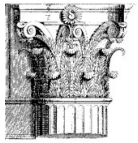
SAN topologies: a small sample





SAN fabric design

- **Current state of the art**
  - designs are done by hand, using a few simple “potted” topologies
- **Surely automation should be straightforward?**
  - Given:
    - flows between endpoints (hosts, devices)
    - link, hub, switch characteristics
  - Apparently not!
    - degree-constraints seems unusual
    - divide-and-conquer seems unhelpful
  - “Extra credit” items are very important:
    - fault tolerance: designing for all possible failure cases
    - multiple layers of switches/hubs possible



SAN fabric design: HPL Appia project

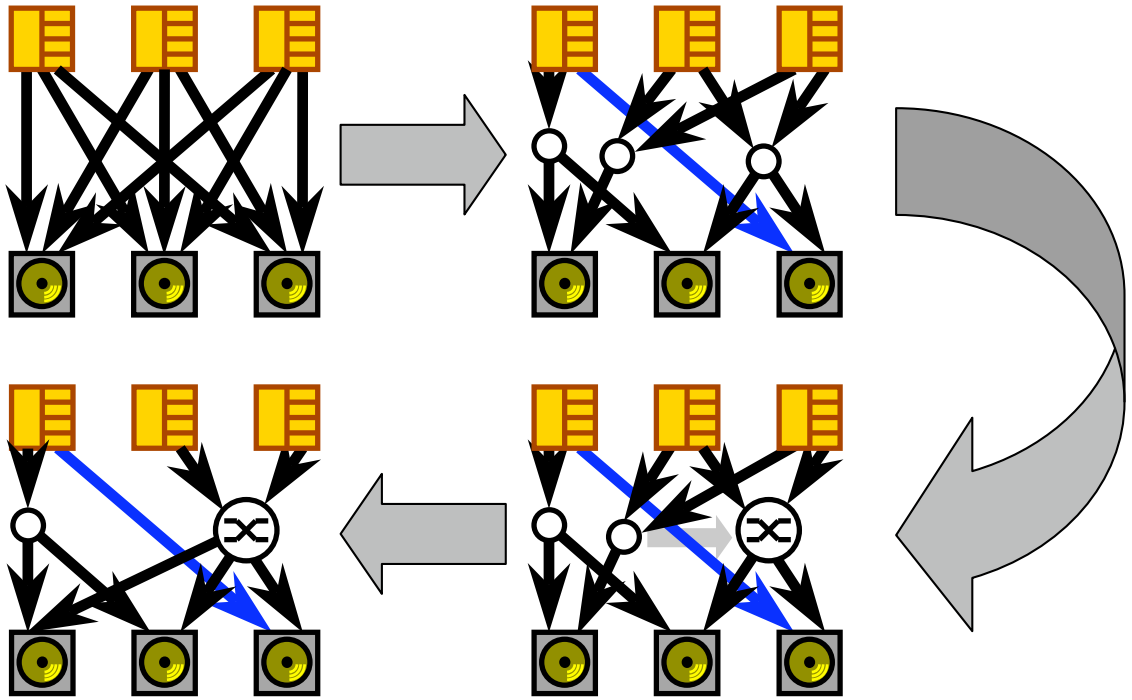
Appia designs FibreChannel fabrics for a Rome assignment

Benefit: error-free, near-minimal-cost fabric designs

This diagram shows the operation of the MERGE heuristic (developed by Li-Shiuan Peh at SSP/HP Labs) as it finds the best solution to a simple SAN design problem.

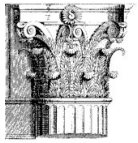
Future work will extend this to design HA fabrics with performance guarantees.

We'd love some help!



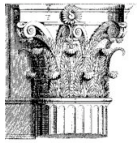
Storage area networks

- Physical connectivity
  - links and loops
- Command sets and protocols
  - what to say
  - how to send it
- LAN vs SAN
  - and other false dichotomies
  - case study: CMU NASD



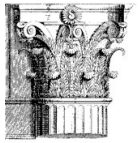
Storage interfaces: SCSI-3 command set

- **Fixed-size *Command Descriptor Block***
  - **read, write, inquire, …**
  - **think of this as an rpc**
- **Separate read- or write data phase**
  - **transfer controlled by the “target” (aka, disk)**
  - **simplifies buffer management for small devices**
- **Mode pages**
  - **data about the device**
  - **setting/reading configuration information**
- **Asynchrony**
  - **multiple outstanding requests**
  - **limited sequencing: “put at front”; “add to end”**

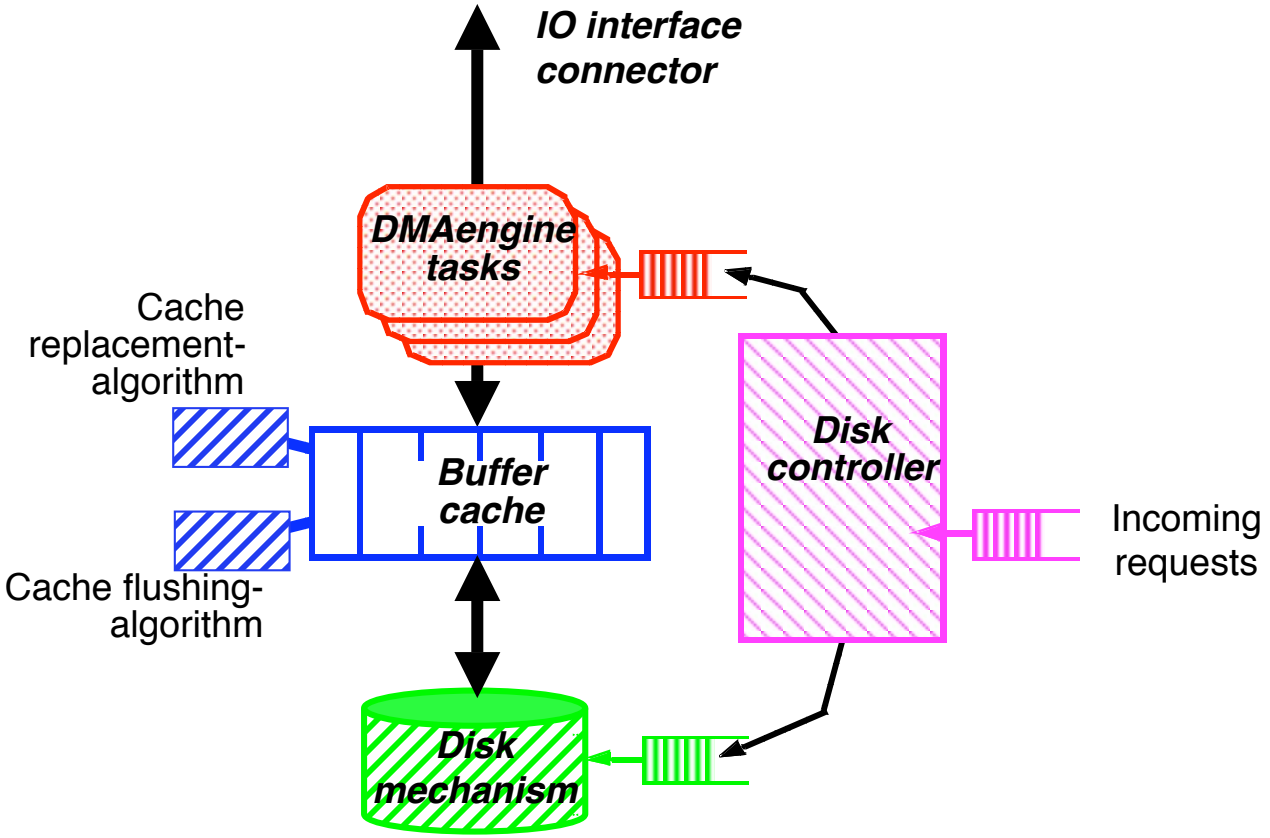


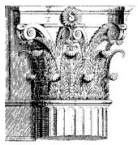
Storage device interfaces - protocols

- **Functions:**
  - buffer management
  - flow control
  - error recovery
  - security
- **Existing choices**
  - parallel SCSI: bus-based signalling
  - FCP: mapping of SCSI signalling onto FibreChannel
- **Future possibilities**
  - a new block-transfer protocol?
  - TCP/IP + RDMA extensions (Cisco, draft of Feb. 2000)
    - advantages: existing management infrastructure; high speed; standards organizations
    - disadvantages: security problems; no multidrop

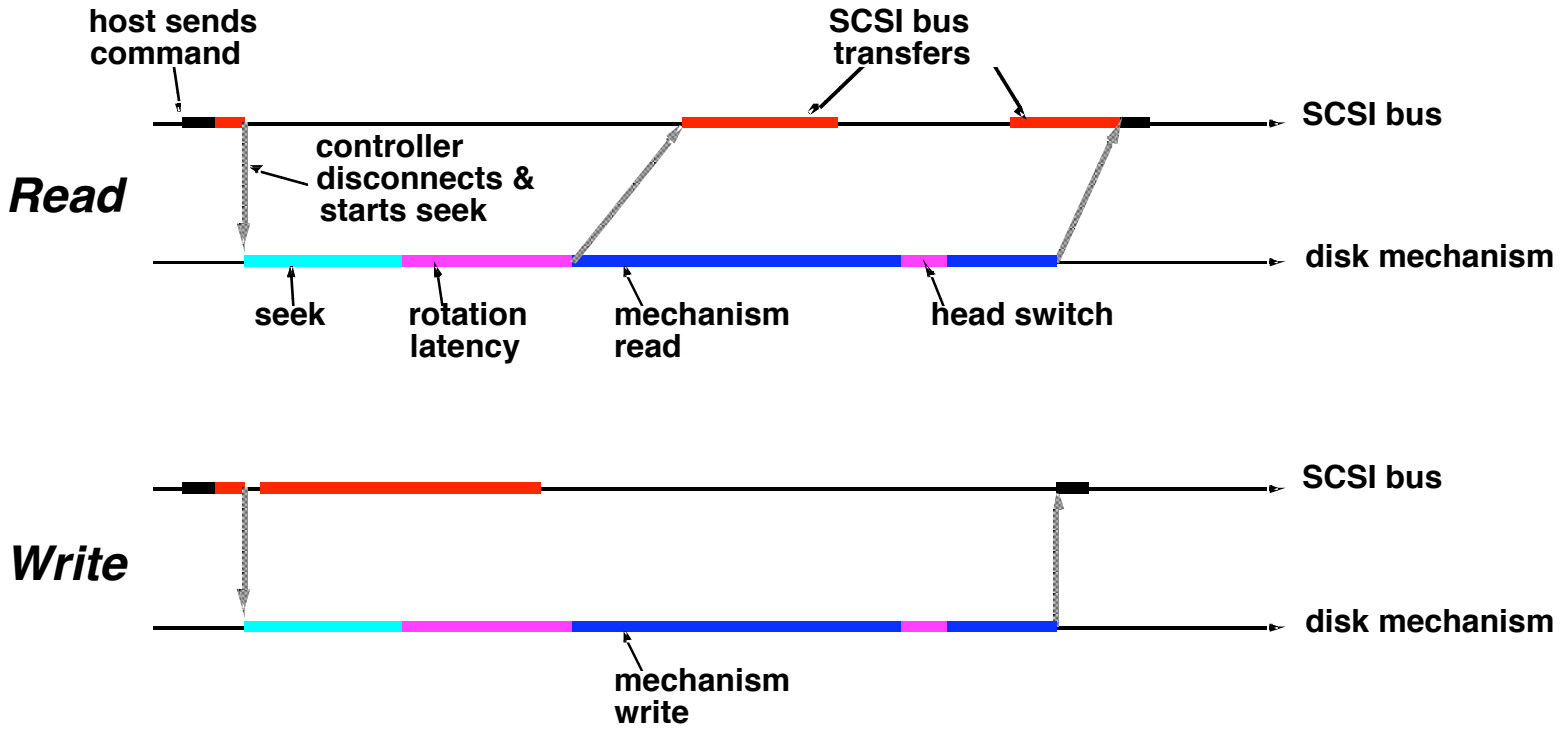


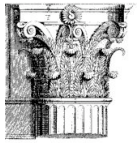
Storage device interfaces: data movement





Storage device interfaces: SCSI read/write

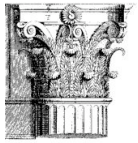




Storage device interfaces: security

- **Security**
  - this wasn't a factor in locally-attached storage
    - now, NT thinks it can format any device it can reach!
  - solutions:
    - zoning (switch vendors) -- pervasive, but coarse granularity
    - host-side security (HP Storage Manager DM) -- defeatable
    - device-side security (EMV VolumeLogix) -- slow to deploy
  - roles are important
    - "can host X see this device?"
    - "can host X read/write to it?"
    - "can host X configure it?"

- **Discovery: what storage is out there?**
  - naming infrastructure
  - multi-path detection

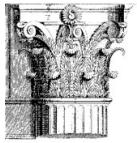


Storage device interfaces: smart storage

- Run some of the application function in the disk (array)
  - Eric Riedel, Kim Keeton, Mustafa Uysal all worked in this area
- Big benefits if:
  - embarrassingly parallel application
  - ratio of data-looked-at to data-shipped-to-host is high
    - e.g.: database select operation in decision support (4x improvements)
    - e.g.: parallel sort
- A few of the open issues
  - programming model
  - resource management (especially with multiple applications)
  - error management/containment/security
  - support

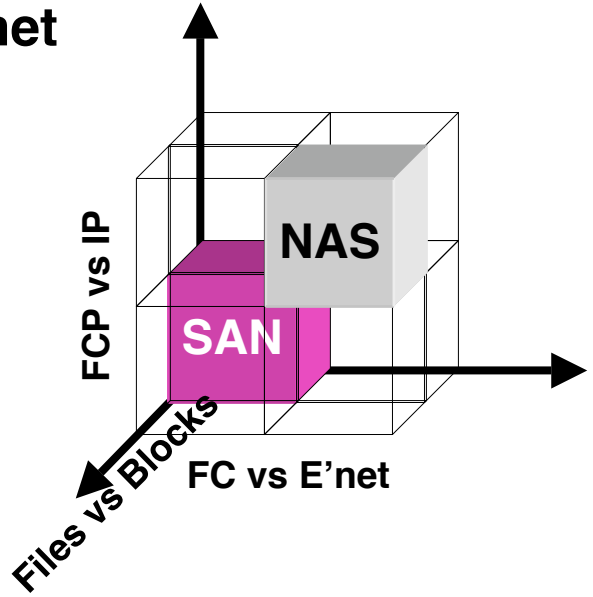
Storage area networks

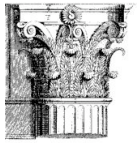
- Physical connectivity
  - links and loops
- Command sets and protocols
  - what to say
  - how to send it
- **LAN vs SAN**
  - **and other false dichotomies**
  - **case study: CMU NASD**



LAN vs SAN vs NAS?

- **Network hardware: FibreChannel vs Ethernet**
  - 1Gb/s available today
  - 10Gb/s E'net will (probably) be ready first
- **Storage interface: blocks vs "files"**
  - block storage devices (SCSI)
  - "NAS" => file servers (Netware, NFS, CIFS)
- **Network protocol: FCP vs TCP/IP**
  - specialized protocol vs general-purpose one
- **SAN:**
  - dedicated network, used (largely) for storage
  - *whatever the protocol!*

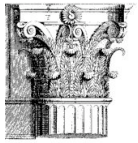




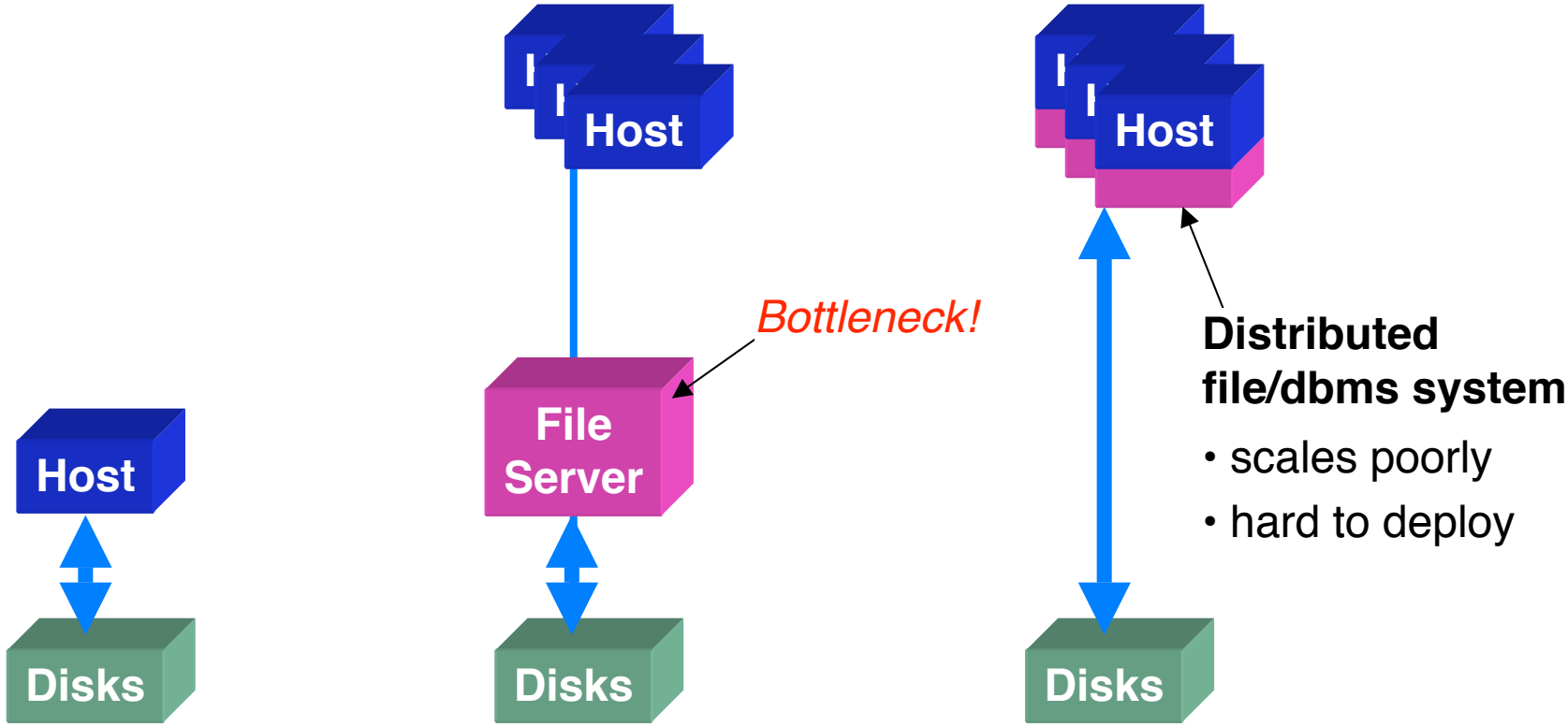
Blocks vs “files”?

- **Block storage devices (SCSI)**
  - **critical path simple => fast**
  - **difficult to push function down to storage device**

- **“NAS” - file servers (Netware, NFS, CIFS)**
  - **can optimize layout and caching for prefetching, readahead, write-behind, etc, …**
  - **finer-grained protection possible**
  - **critical path has another layer of mapping => slower**



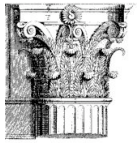
Blocks versus files? Some choices



Locally attached storage

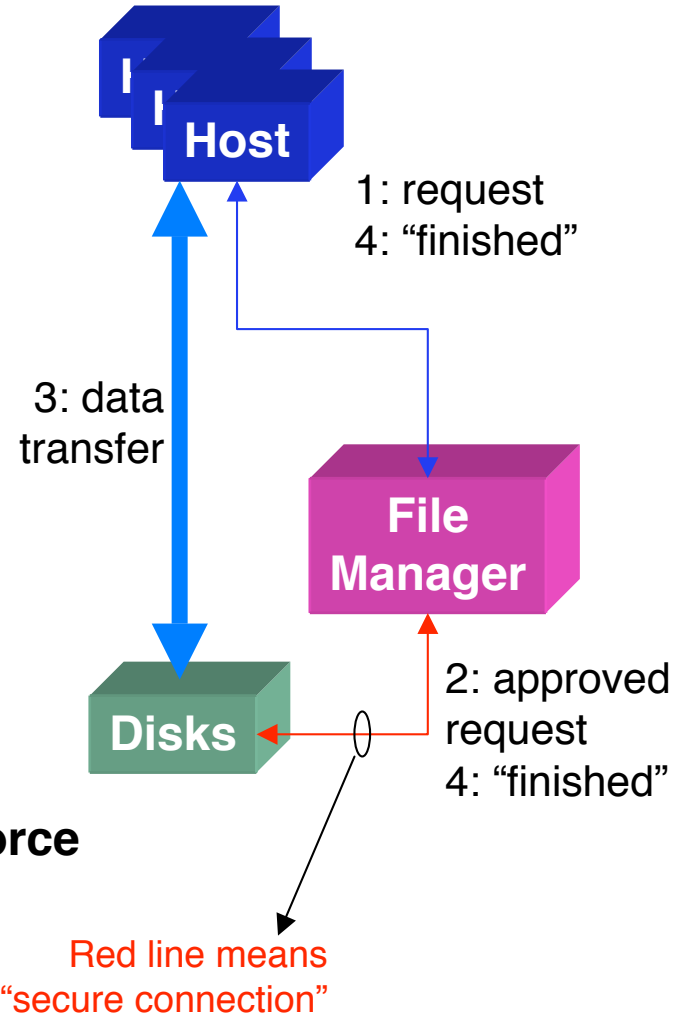
File server attached storage (CMU "SAD")

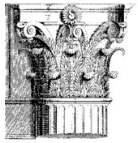
Shared storage pool plus distributed FS



Blocks versus files? CMU "NetSCSI"

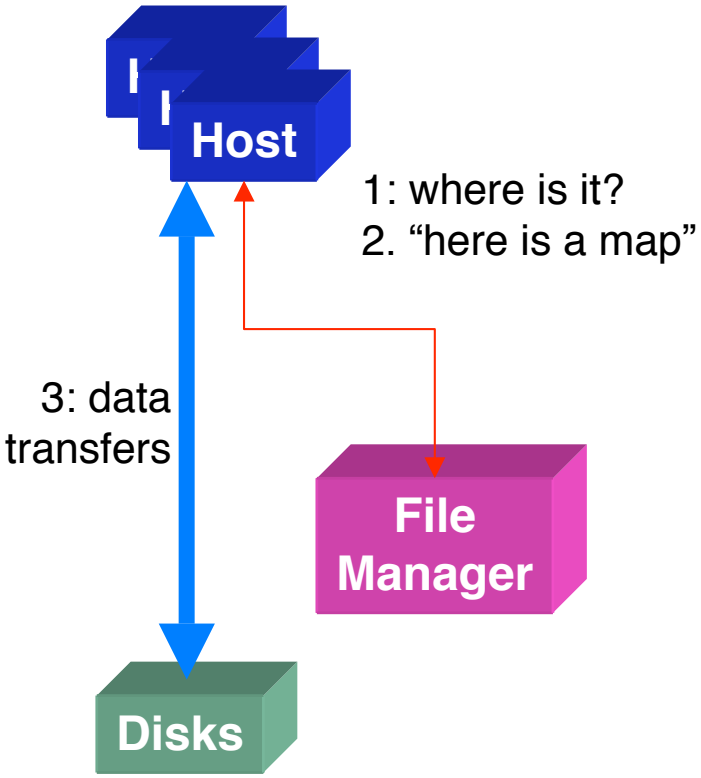
- **NetSCSI**
  - **use a "file manager" to police the requests and manage the data layout**
  - **data flows directly to/from the host**
  - **otherwise like NFS**
- **Advantages:**
  - **great for large data transfers**
  - **fine-grained protection**
- **Disadvantages:**
  - **file manager is still bottleneck**
  - **requires secure channel to disk to enforce protection rules**

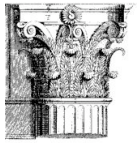




Blocks versus files? Transoft DFS

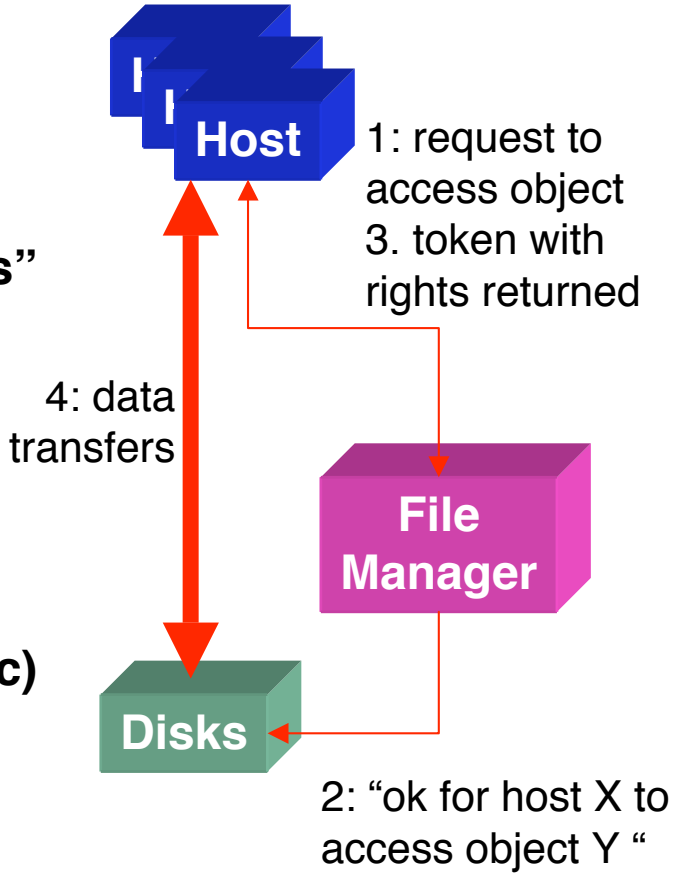
- **Transoft DFS (CIFS-based)**
  - **use a "file manager" to police the requests, and manage the data layout**
  - **"layout map" returned to host**
  - **data portion of transaction flows directly to/from the host**
- **Advantages:**
  - **file manager can be CIFS server**
  - **great for large data transfers, repeated transactions**
  - **existing SAN infrastructure for IOs**
- **Disadvantages:**
  - **protection granularity is whole LUN**
  - **file manager may still be a bottleneck for metadata *changes***

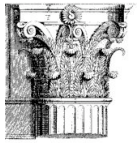




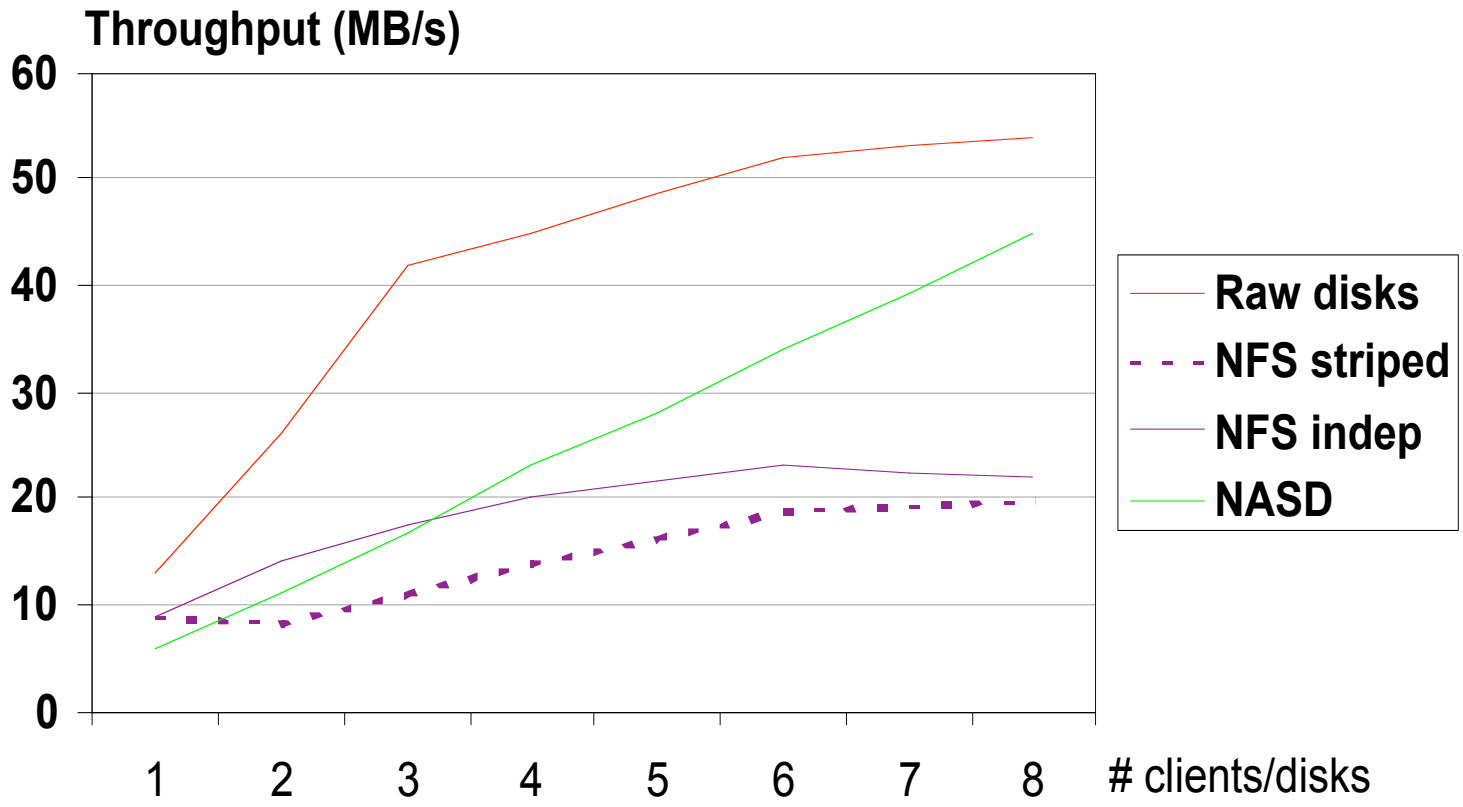
Blocks versus files? CMU NASD

- **CMU NASD**
  - **use a "file manager" to police the requests, and manage the data layout**
  - **"permission token" returned to host**
  - **expressed in terms of "storage objects"**
  - **rest of transaction flows directly to/from the host**
- **Advantages:**
  - **best performance for NFS-like loads for workgroups (lots of metadata traffic)**
- **Disadvantages:**
  - **fine-grained protection requires a *lot* of work on device**
    - **CMU: requires file system in disk**
    - **ISI NetStation: uses dynamic map instead**





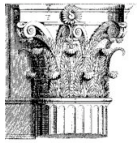
CMU NASD: sample performance results



133MHz Alpha NASDs; 233MHz clients; switched 155Mb/s ATM SAN
500MHz Alpha NFS server; dual 155Mb/s ATM links
Finding parallel association rule on 300MB of sales records

Carnegie Mellon

HEWLETT PACKARD



CMU NASD: what about the cpu cycles?

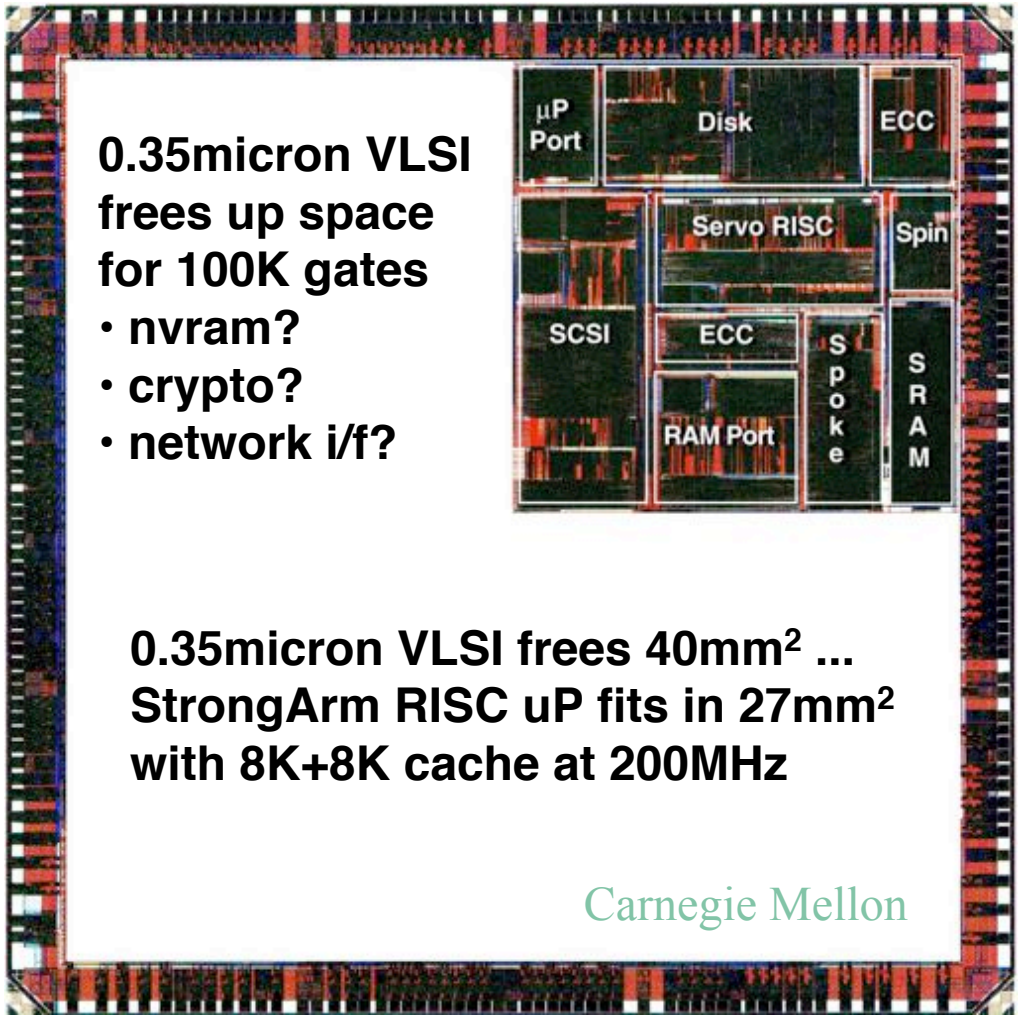
Quantum Trident drive

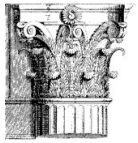
"Today" (1997): M68020 + datapath ASIC (at right)

- **0.68 micron (74mm²)**
- **4 clock domains, each 40 MHz:**
  - SCSI processor
  - disk R/W channel
  - uP control port
  - DRAM port
- **~110 Kgates + 22Kb**

Since then:

- **Siemens TriCore**
- **Cirrus Logic 3CI**
- **TI TMS320C27x**

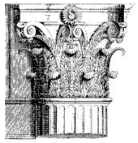




CMU NASD model: a personal opinion

Some of the good results:

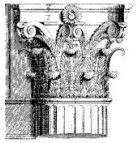
- **Security model [Howard Gobioff]**
  - **much more resilient to attack than most SANs**
  - **shows how to provide fine-grained device sharing**
  - **precompupted digsts speed cryptography**
- **Object model**
  - **basis for protection, pre-fetching, layout, ANSI standard proposal**
  - **offloading NFS operations from file manager can help**
- **Framework for smart storage devices**
  - **security + object model**
  - **“Active disks” combines nicely with the object model**



CMU NASD model: a personal opinion

Open question: is it the right answer?

- **Probably not such a good a match for:**
  - **high end databases** (dbms tables are larger than devices)
  - **low end desktop** (no desire for file system in drive)
- **Is potentially a good match for scalable mid-range file service**
  - **eg, IDC/ASP environment**
- **Book is still open on NASD vs Transoft DFS models**
  - **performance strongly affected by workloads**
- **Lots of good ideas/technology have resulted**



SANs -- conclusions

- SANs are an important enabler for the storage-utility model
  - they permit rapid growth and resource redeployment
- SAN vs LAN is the wrong question :-)
  - 3 independent decisions:
    - link technology
    - command-set
    - protocols
  - But … LAN technology will probably sweep away FibreChannel-based SANs in the next few years
  - TCP+RDMA/IP seems a strong contender
- Smart devices will change the landscape
  - when?
  - security may prove a decisive factor