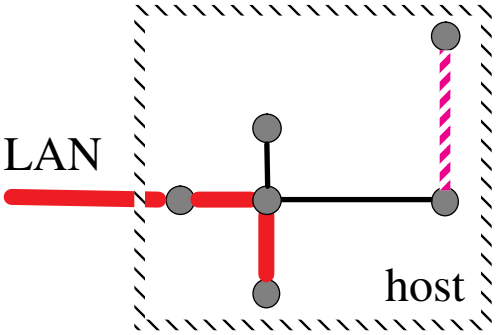
Summary: Netstation Properties

1. Physical Attributes - Gigabit channels.

   No slot limit. Bandwidth scales.

   Hardware routing and channel performance tracks VLSI progress.

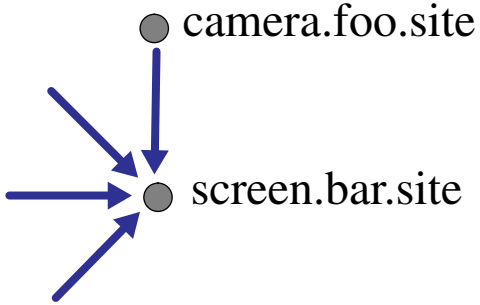


2. Logically Interfaced → Interoperability

   Protocols create a logical rather than physical interface.

3. Separable and Composable

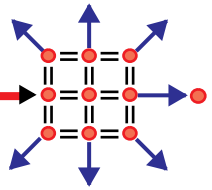
   NVDs tied to network, not a chassis.



4. Internetwork Addressable

   They are hosts.

   NVDs can be destination of multiple streams.



Research Implications

Standards: Protocols

Each class of device requires a logical interface.

- **Must define protocols for displays, disks, processors, . . .**
- **Host explosion . . . the single–chip host is already here.**

Security: Visibility

Sanctity of the chassis is gone.

- **Must acquire and retain control over one's devices.**

Routing: Multiply Addressed

Location becomes a soft concept.

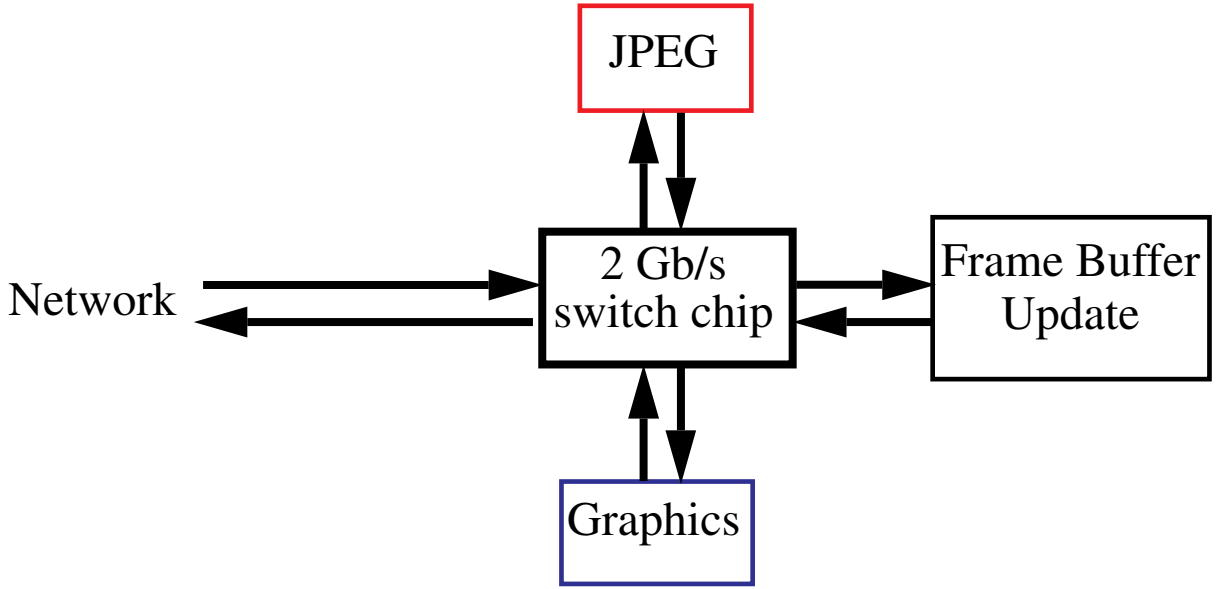
- **Workstation is a set of addresses and routes.**

Speculation: As the VLSI Turns

A thin-node gigabit interface already fits on the corner of a chip.

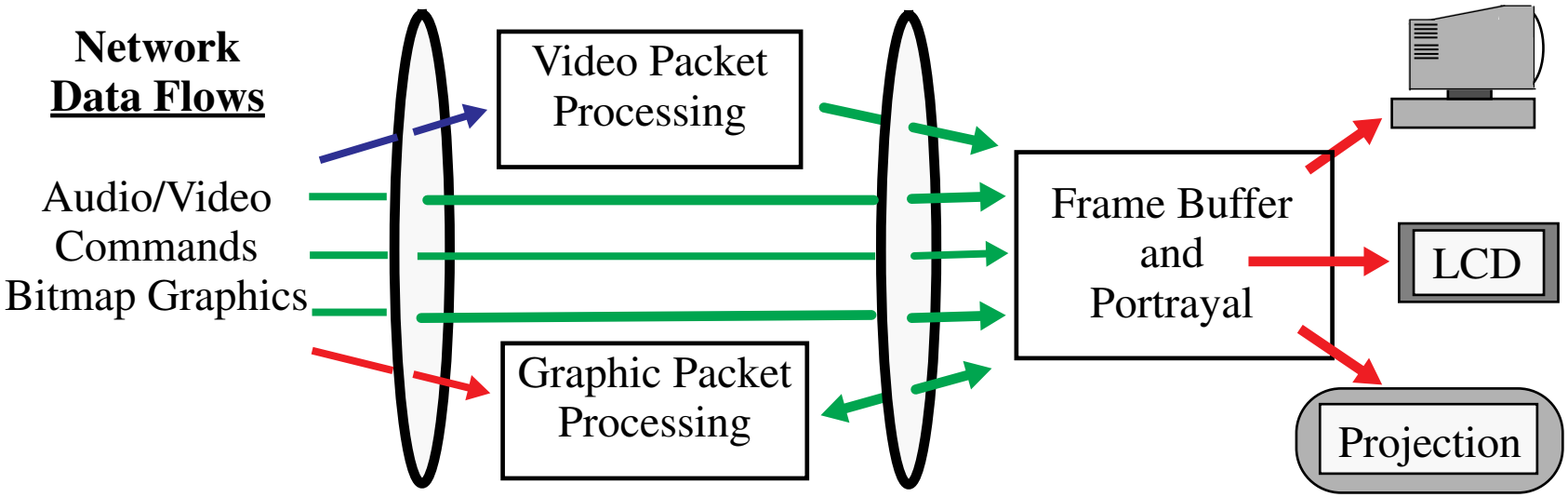
- **It will only become smaller and faster . . .**
- **Can it become an ASIC cell - maybe?**

Packets sent to specialized 'hosts' are transformed into rasterop packets.
Those in turn are sent to update display memory.



Rasterop packets might be generated for the display virtually anywhere.

Logical Interface Points

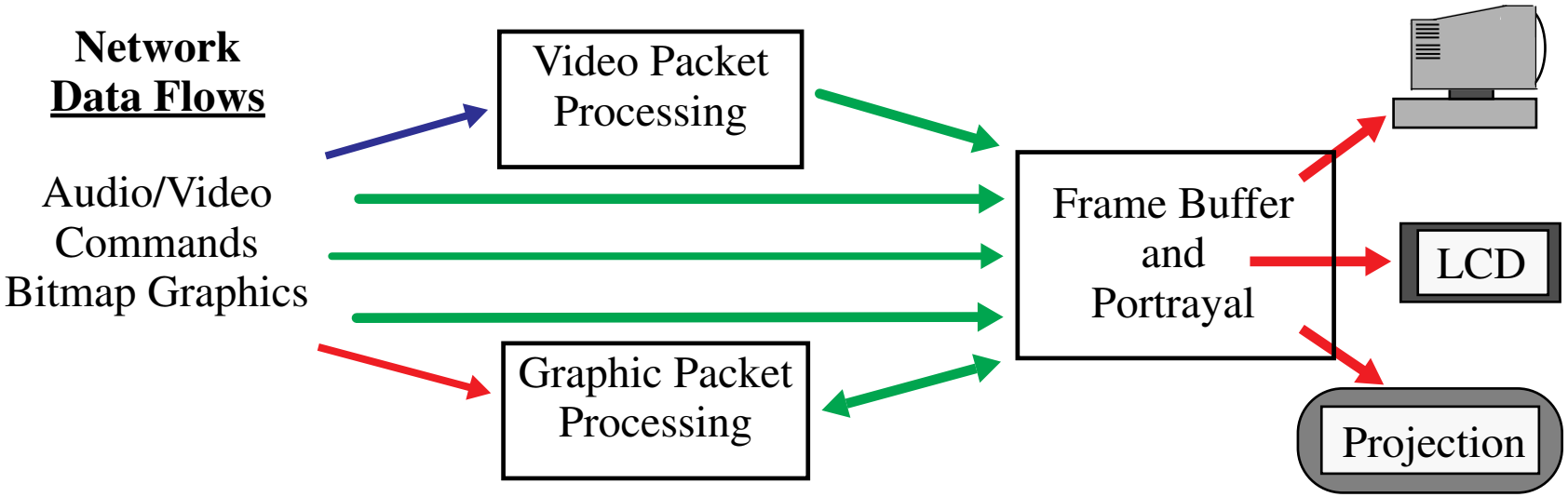


There are two realistic interface points.
The leftmost produces a more conventional display controller.

- **Emulate a 'frame buffer' to allow bitmap graphics.**
- **Video - JPEG, MPEG-II, or possibly uncompressed.**
- **Command - Provide for device control and security.**

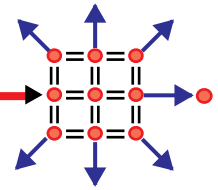
Display NVD

Many types of data flows arrive at a display.



Need to define a set of suitable protocols at an interface point.

Where do you make that cut?



Network Virtual Device - (NVD)

If a device attaches only to the network ---

It interacts only via protocols ---

It is a virtual device.

Networks + protocol standards ➔ interoperability.

- **Networking's greatest benefit (Do you use FTP and Email?)**

NVD Definition Procedure

Capture device control and function logically

. . . in terms of distinct data flows to and from the device.

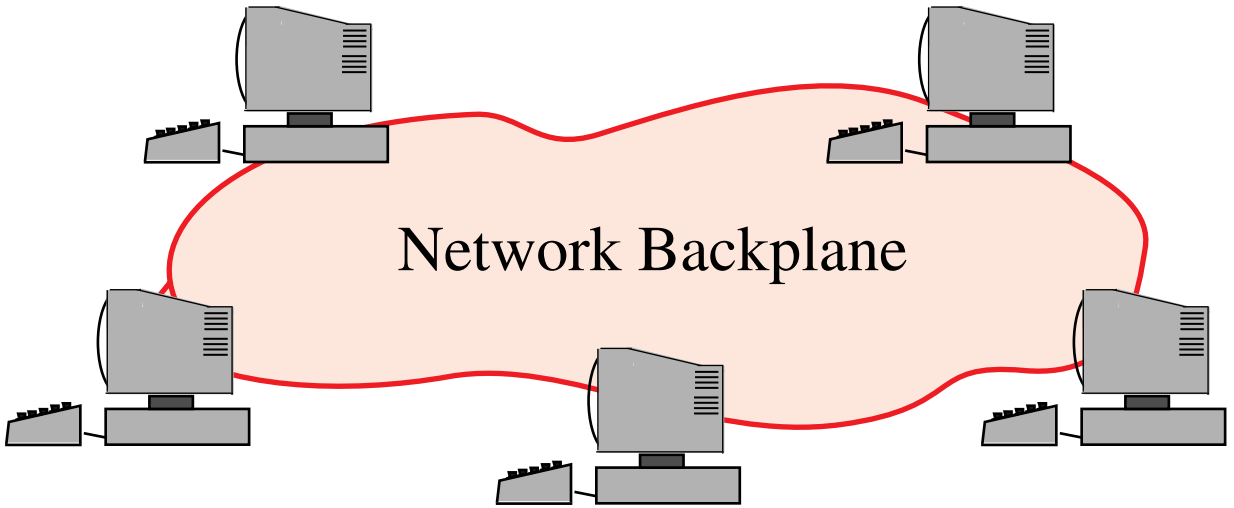
- **Define protocols suited to each data flow, capturing its nature: real-time, reliable transport, datagram . . .**

Cluster Computing Support

Reliance on thin-node technology places support for Gigabit message transmission in hardware.

→ Parallel computing support is 'built in'.

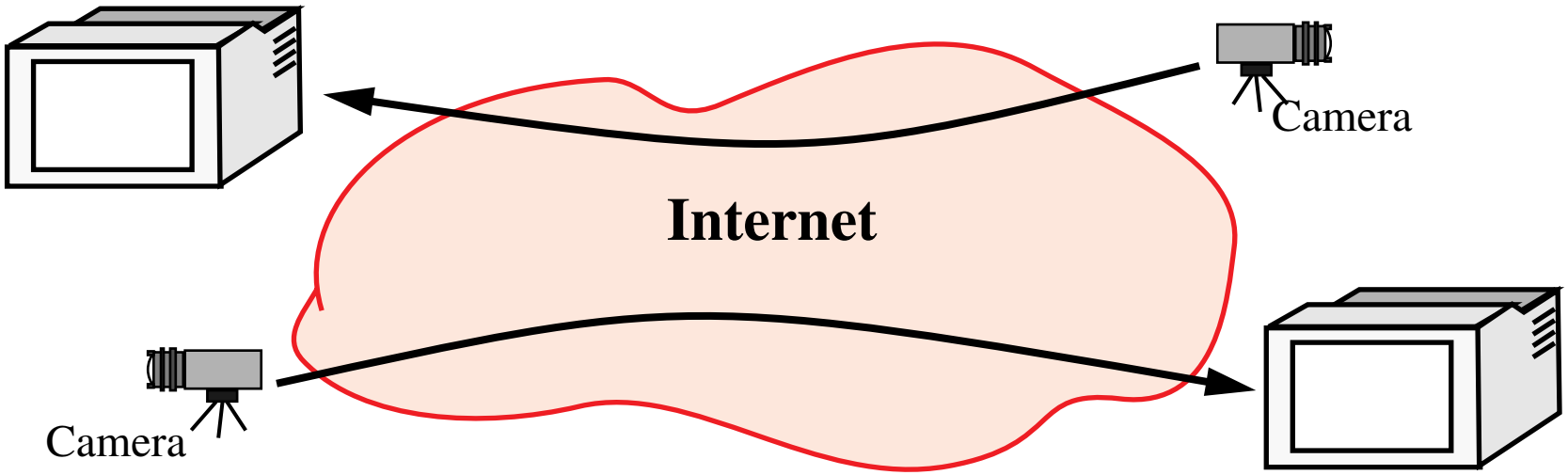
- **Little difference between cluster of workstations and a multicomputer.**
- **Workstations are "fat-node" parts of a parallel supercomputer.**



All Nodes Have Equal Access

Nodes have direct access to the Gigabit LAN and Internet.

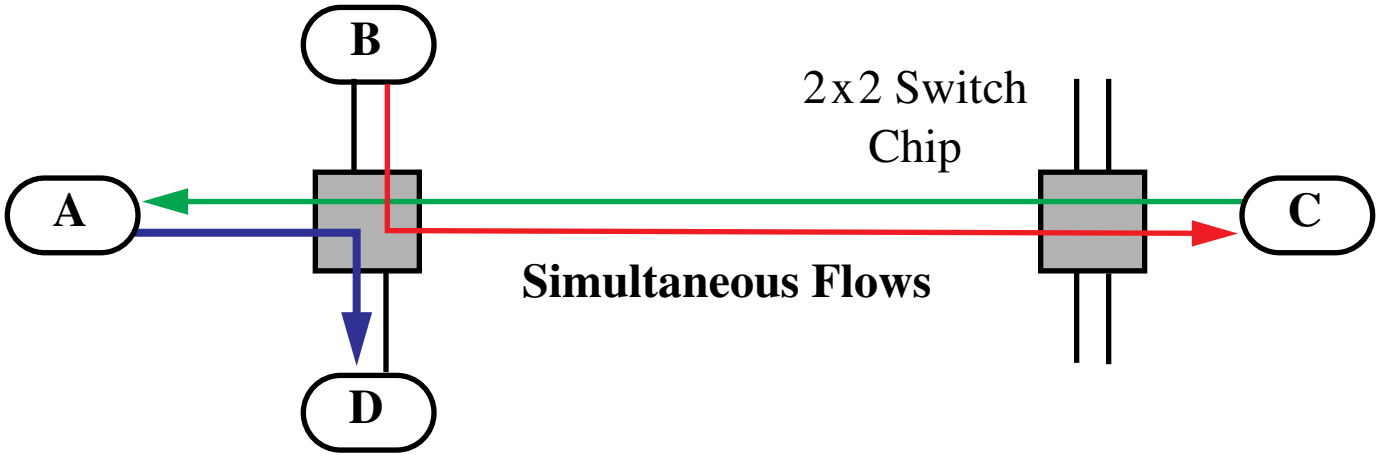
- **Packets pass without requiring help of the workstation CPU.**
- **Relies more upon routing hardware than software.**
- **Direct support for "conferencing" applications.**



4. Consequences: Software and Hardware

System Bus Becomes Superfluous

DMA . . . is replaced by . . . point-to-point packet transmission.

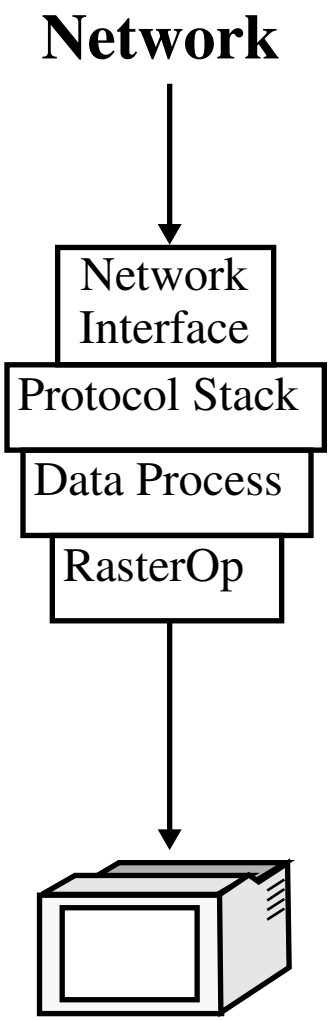


Enhanced Flexibility

Slot-count limitation --- gone. **Want another device --- add a new host.**

Bus bandwidth ceiling --- gone. **Network capacity scales with growth.**

Artifacts of Centralization Removed



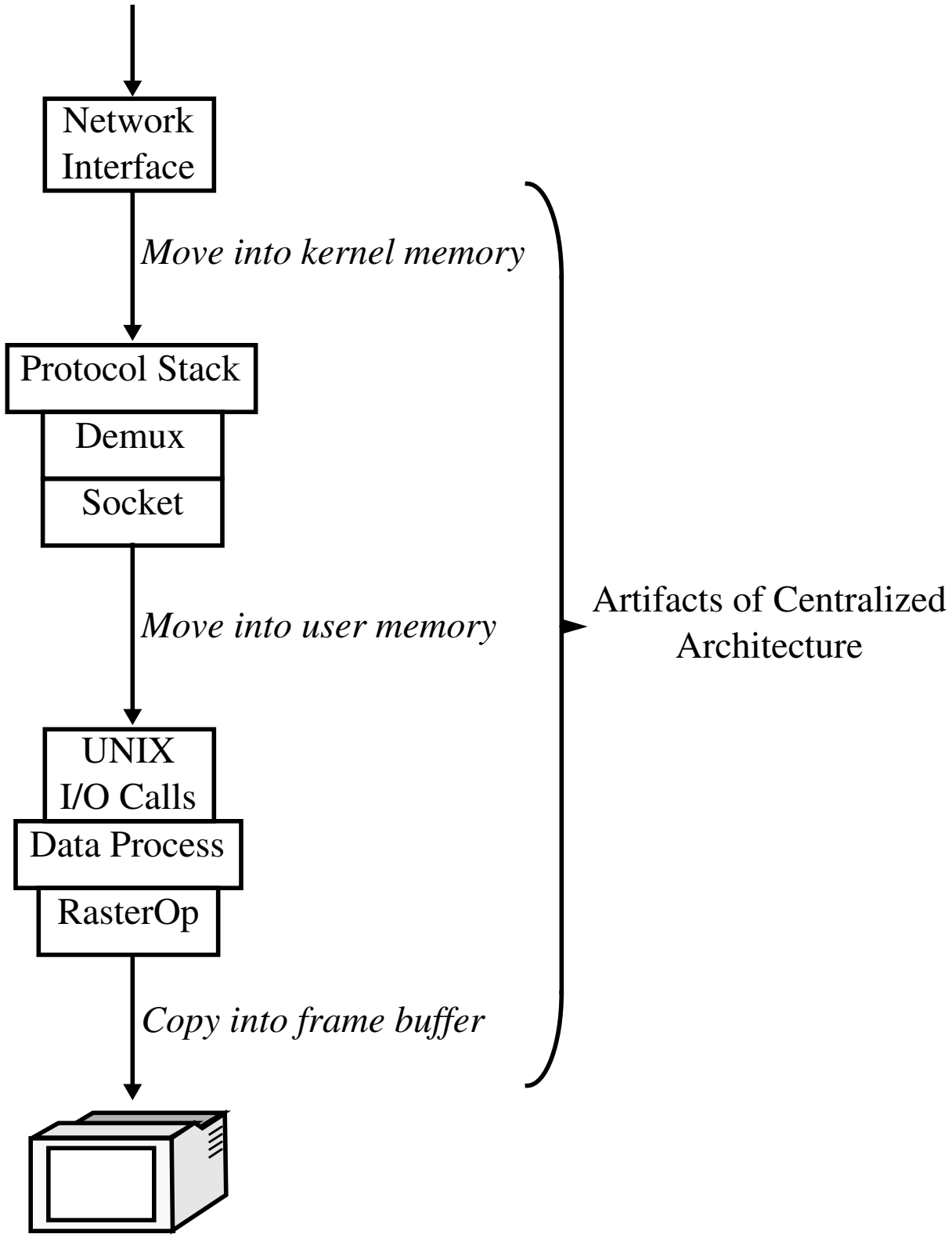
Display is accessed via its own gigabit LAN interface.

Consequences

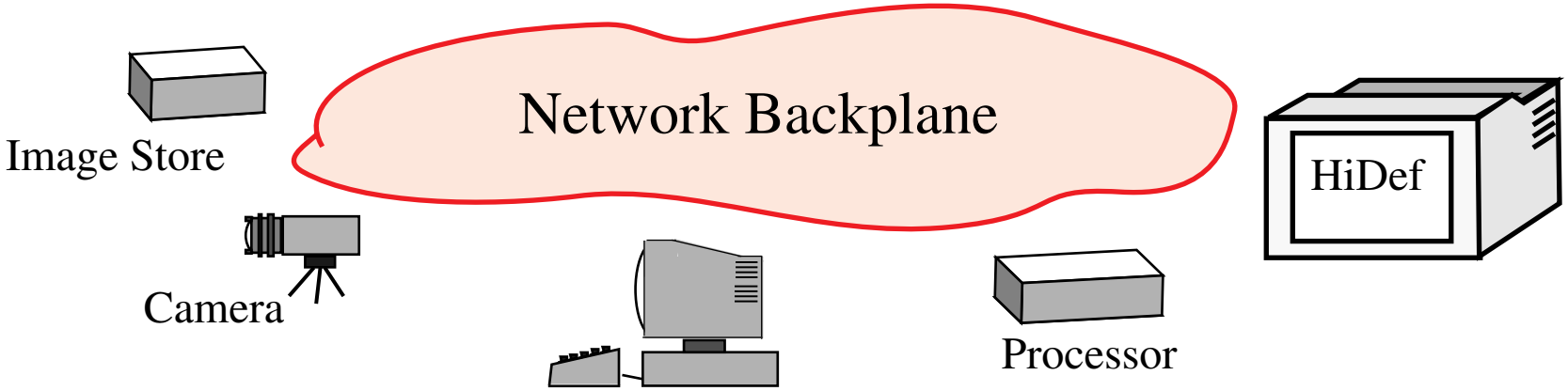
1. **Display has its own Internet address. Packets are routed directly to/from it.**
2. **Workstation OS no longer acts as an intermediary.**
3. **OS Socket, & I/O call overhead ---- gone. Data movement is minimized.**

Receiving Display Data Packets

BSD UNIX

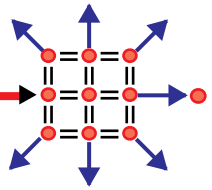


Netstation Architecture



Give processors and major devices thin-node LAN interfaces.

This results in a network-centered architecture . . . or a huge distributed parallel computer . . . depending upon your viewpoint.



Workstation Architecture

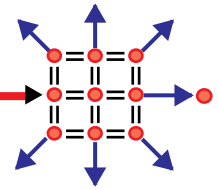
Orthodox architecture is centralized. It shares one network interface. DMA achieves One-Copy operation.

- **Single network interface → N:1 relationship on output, 1:N on input.**
- **Protocol stack is centralized and resides inside OS. Packet Mux and DeMux is done by software.**

Unorthodox Architecture

Decentralize. Use thin-node interfaces. Bring the local area network 'closer' to each processor or major device.

- **Produces a 1:1 relationship for both output and input.**
- **Protocol stack distributed. DeMux done by routing hardware.**



RPC Transmit Performance--- SPARCstation-10

	RPC/s
UNIX stack	**2000**
Zero-Copy ALF	**20,000**

User IP/UDP sent loopback (not out the network).

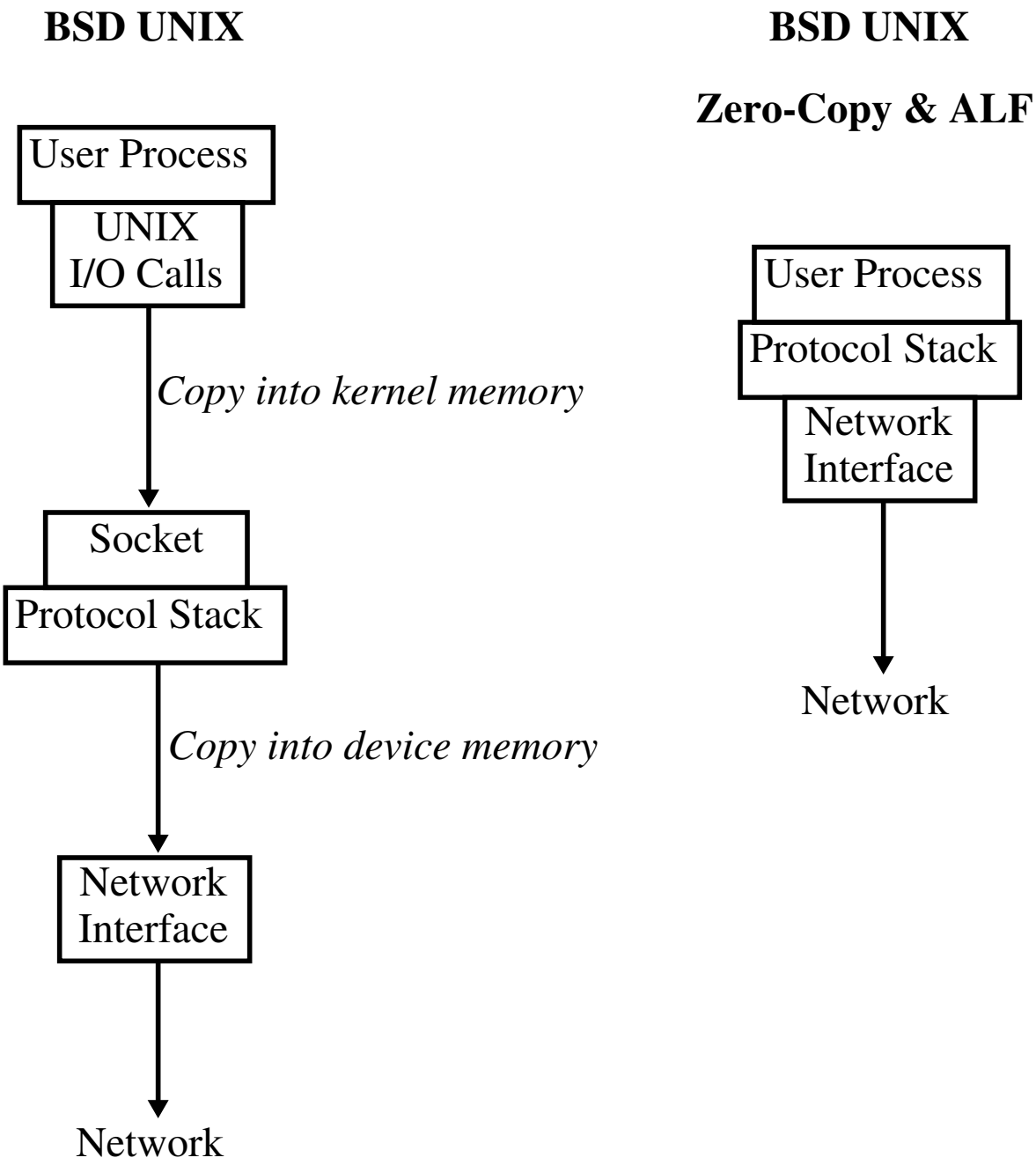
User to User transmitted across ATOMIC, using application layer framing and zero copy software.

Zero Copy + ALF produced a 900% improvement.

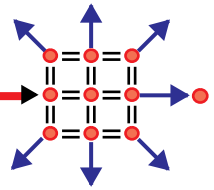
Minimal data movement	✔	**No DMA** **No progrmmed I/O**
Minimal packet processing	✔	**No I/O calls**

Moving the network ‘closer’ to the application works very well.

Sending RPC Packets



How much improvement?



How Good Is Zero Copy and ALF?

Consider remote procedure calls (RPCs).

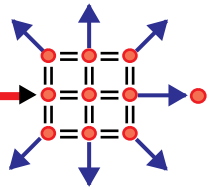
- **Allocate/map per RPC user buffers in the network interface.**
- **User puts RPC packet stencil in a buffer, using it when sending RPCs.**
- **User looks in other buffers for RPC replies.**

IP + UDP + Sun RPC packet -- X(a,b,c) -- ~ 96 bytes.

For each RPC sent the user changes:

Destination address (possibly)
Packet length
Incremental checksum
Procedure, program index and arguments

- sets a go flag - interface sends - resets flag - no system I/O calls



Improving Network Performance

Gigabit goals are met when:

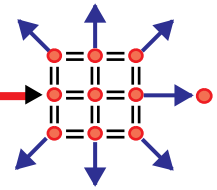
Zero-Copy

Application Layer Framing (ALF) **(Clark, Tennenhouse)**

Orthodox Path

Partially meet goals by:

Not using programmed IO	**- DMA packets**
Not checksumming in software	**- Make part of DMA in net interface**
Not copying into destination	**- Page align and VM remap packets**



3. Motivation and Design of a New Workstation Architecture

Achieving Gigabit performance is not easy. Why?

Performance Bottlenecks

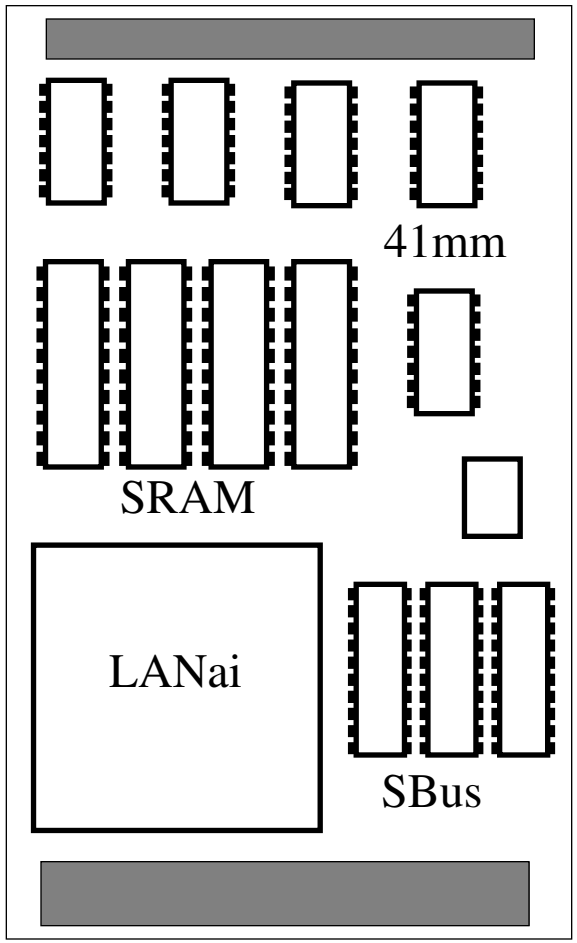
Hardware Architecture	-	System buses
Software Architecture	-	Protocol stack and OS
Packet Processing	-	Too costly

Gigabit Performance Rules

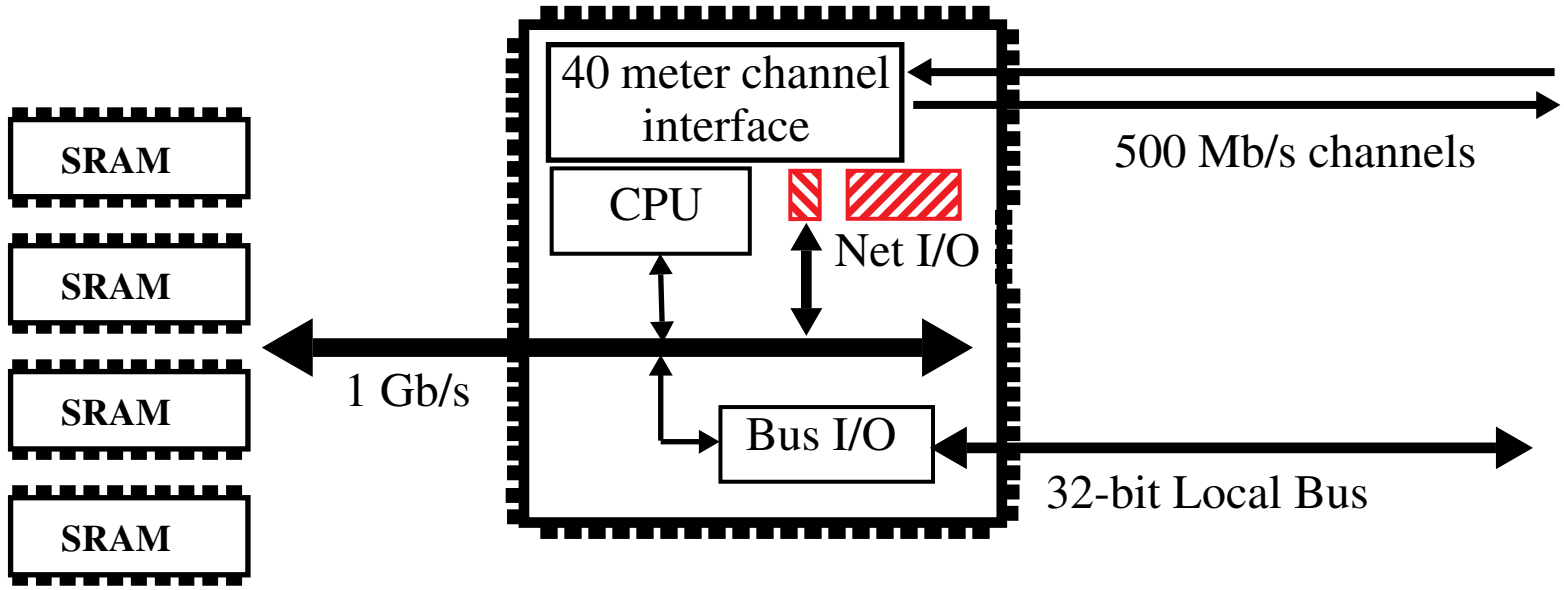
Minimize data movement.

Minimize per-packet processing.

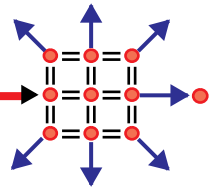
SBus Interface



Mosaic Successor - LANai



	32-bit Transfer	Signal count
VL–Bus:	30 ns	92
SBus:	40 ns	82
Turbochannel:	40 ns	44
ATOMIC:	*66 ns* (per-channel)	*18* (channel-pair)



ATOMIC Performance

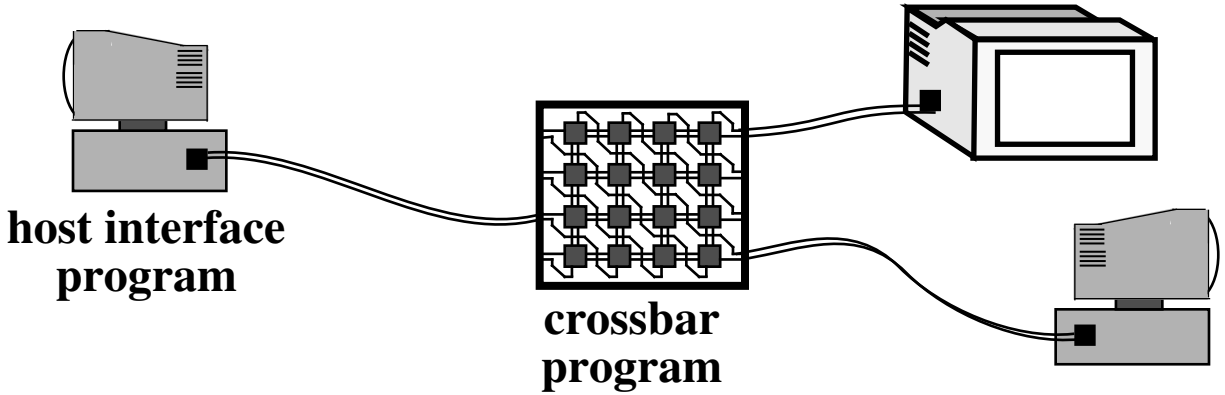
Byte/pkt	Kpkt/s	Mb/s
4	657	21
54	396	171
1500	31	381

Single source. SRAM limited to 400 Mb/s.

Byte/pkt	Kpkt/s	Mb/s	Flows	
2	5250	84	8	Below capacity
54	793	343	2	
1500	33	405	2	At capacity

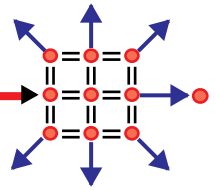
Multiple sources competing for single path. SRAMs limited to 400 Mb/s.

A Thin-Node LAN - ATOMIC



1. **Distribute Mosaic chips into hosts.**
2. **Aggregate other Mosaic chips to form switches.**
3. **Develop suitable cabling [AT&T 41MM differential pairs]**
4. **Program the assembly to emulate a LAN.**

ATOMIC is a physically distributed parallel computer that is programmed to be a LAN.



2. Local Area Network

Thin-node combines computation with Gb/s routing capability.
One can tailor-make a "network" with it.

For example:

- **Begin with a datagram LAN.**

  Emulate Ethernet-style interface IDs, broadcast & multicast

  Perform ID → source route translation

  → "Ethernet" compatible LAN that runs at 500 Mb/s.

- **Recognize RPC packets.**

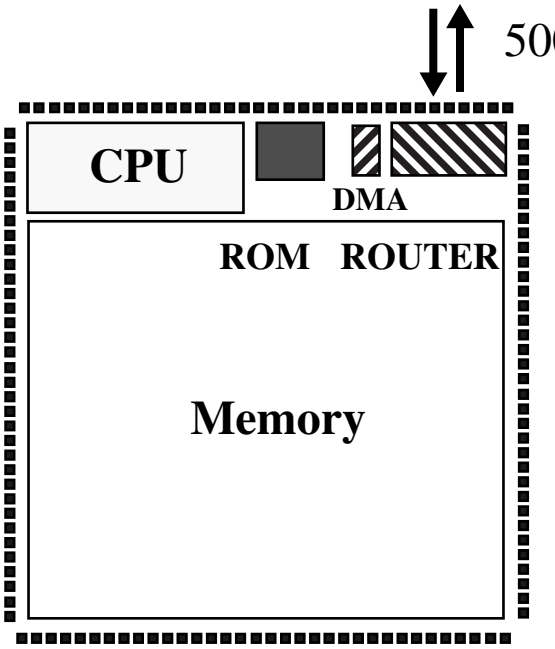
  Source chip assigns Seq# and sends, destination chip ACKs when good.

  Source chip retransmits RPC when ACK not received within Δt.

  → Reliable message service for RPC and Network Control.

Thin-Node Characteristics

1. Sub-chip size gigabit interface.
2. Variable-length messages --- Direct datagram compatibility.
3. Programmable.
4. Scales with VLSI.



Thin-Node Characteristic

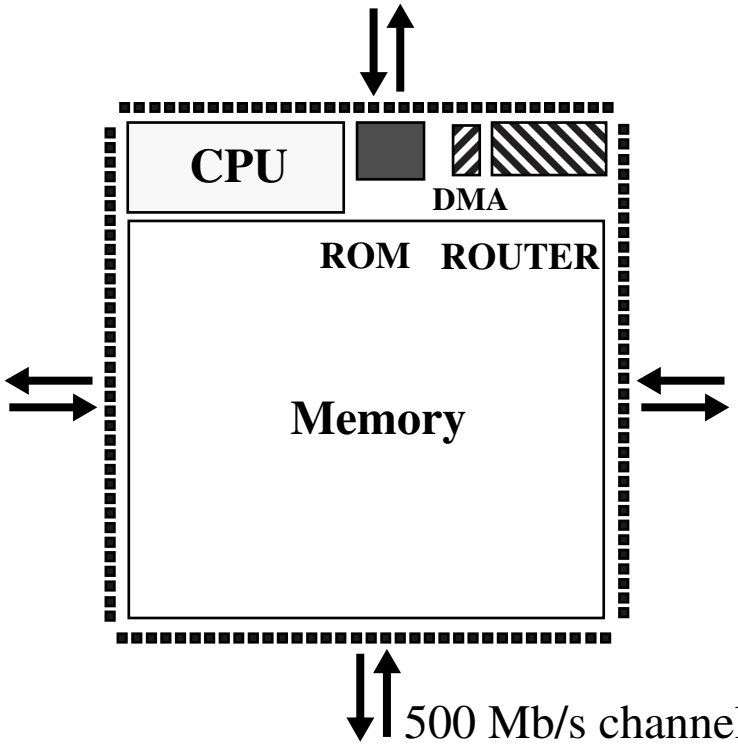
Small interface size.

Direct gigabit interface to your µproc?

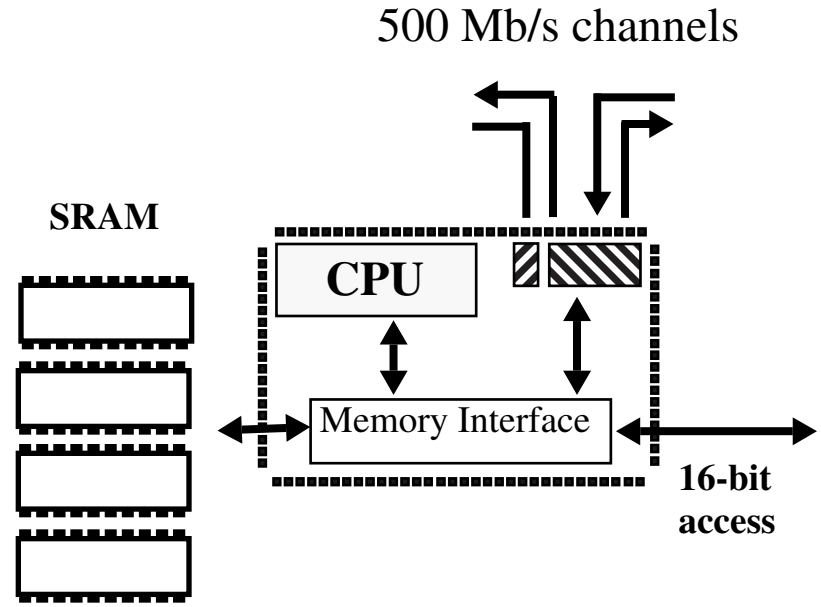
Mosaic Chips

Switch Chip

Network routing and DMA account for 8% of the chip logic.



Interface Chip



Networking and Parallel Supercomputing

Thin-Node

small µproc
small memory
one-few chips
cheap/node
high parallelism

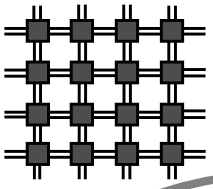
Fat-Node

powerful µproc
large memory
many chips
costly/node
moderate parallelism

Internode Communication Families

Message Passing

Caltech Cosmic Cube



Fat-Node

Hypercube

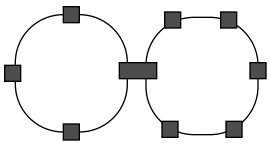
iPSC
Paragon

Thin-Node

Caltech Mosaic

variable-length
mesh topology
source routed
point-to-point
0.5 to 1 Gb/s

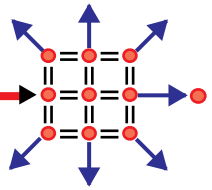
Shared or Distributed Memory



SCI - IEEE 1596 - SCIzzle
Scalable Coherent Interface

discrete-length
ring topology
routed by ID#
1 to 3 Gb/s

+ Thin-node = “interesting”



1. Gigabit Architecture - Where is it now?

Contrary to what we hear . . . 100 or 155 Mb/s is not a Gigabit.

At a Gigabit things get hard.

- **Instruction time consumes a word of bandwidth**
- **Network operates at DMA rate**

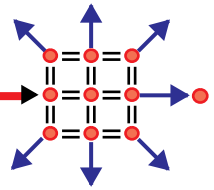
But Gigabit Architectures and Networks Do Exist

Parallel supercomputers have used Gigabit networking for years.

- What has that community been doing?
- Can we use this technology?

Is it practical?

Is it cost effective?



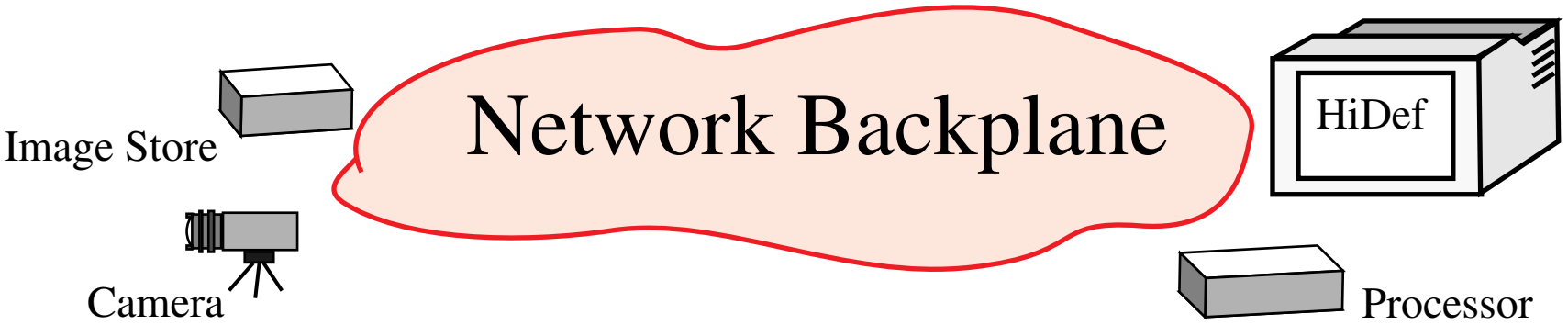
Building Gigabit Workstations - One Designer's Viewpoint

Guiding Principle

Better design envelope is obtained by decentralized design based upon gigabit networking rather than a system bus.

I will present:

1. Developments in parallel supercomputing that scaled gigabit datagram networking down to chip level.
2. Performance of a Gb/s LAN built from that technology.
3. Why it is difficult to obtain Gb/s performance from today's workstation hardware and software architectures.
4. Solution by rearchitecting around Gb/s network. Pros and cons . . . and a possible future.



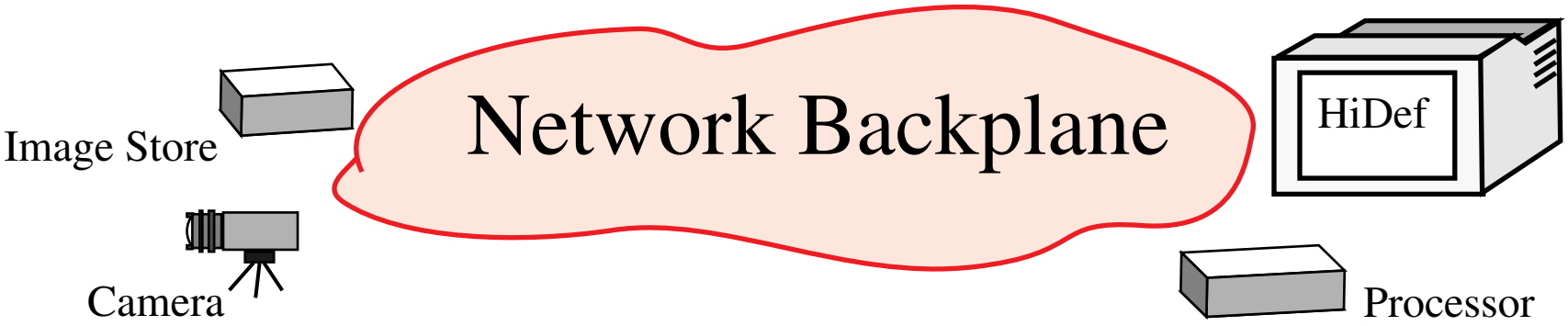
Gigabit Architecture

Workstations and applications are increasingly network-oriented.

- Email, Network Servers, Teleconferencing, MultiMedia

Applications are pushing development of "Gigabit" technologies

- Medical image transfer, Scientific visualization
- Virtual reality, "Information Wonder-Highway"
- JPEG (**NTSC**), HDTV (**MPEG-II**)



Networld + Interop-94

Netstation Architecture
Gigabit Communication Fabric

G. G. Finn
USC/Information Sciences Institute
finn@isi.edu