On the elusive benefits of protocol offload

Piyush Shivam, Jeff Chase

Duke University

Ethernet/IP in SANs

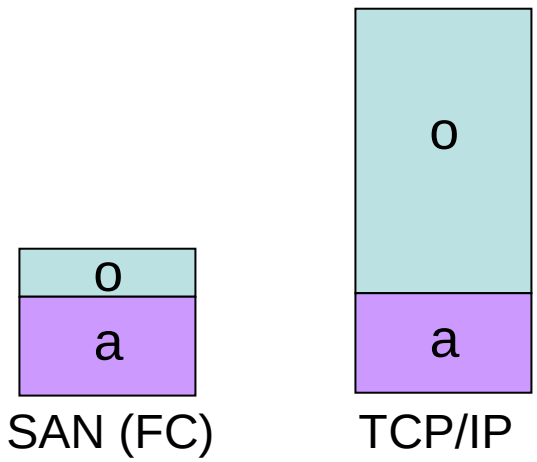
- GigE and 10GigE enable deployment of commodity Ethernet
  - Data center
  - SANs
- Ethernet means IP
- IP as a competitor to ‘specialty’ nets
  - FiberChannel, Infiniband

IP vs. FC, Infiniband etc.

- Advantages of IP
  - Cost effective
  - Standards based
- Questions
  - Can IP really compete?
  - What are its problems?

High host overhead

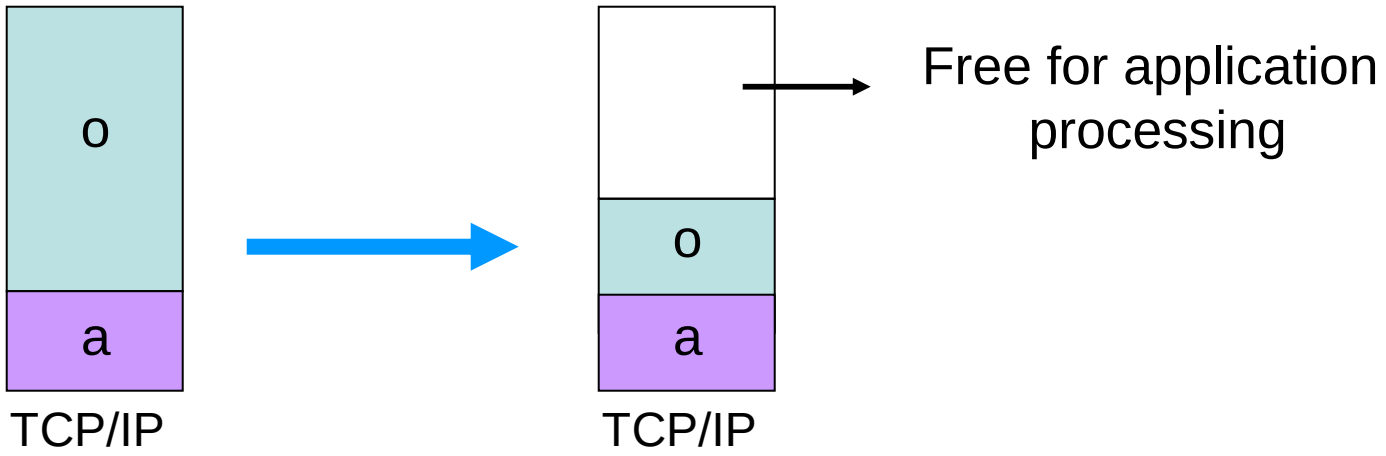
- TCP/IP processing
  - Managing the dumb NICs
  - Takes away CPU/Memory cycles from the application



a = application processing per unit of bandwidth
o = host communication overhead per unit of bandwidth

Solution

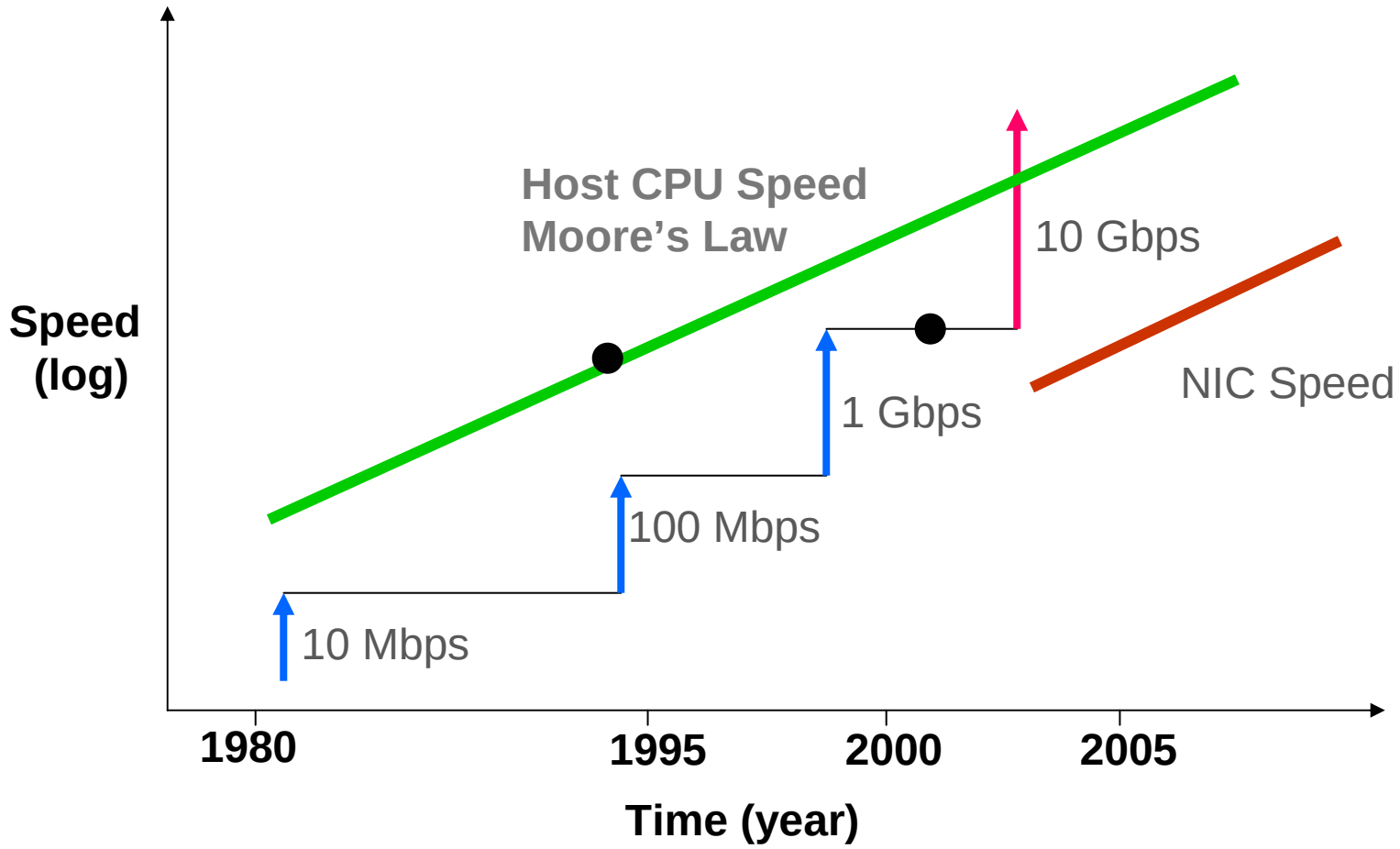
- Smarter NICs
  - Offload TCP/IP processing from host
  - Free host cycles for the application



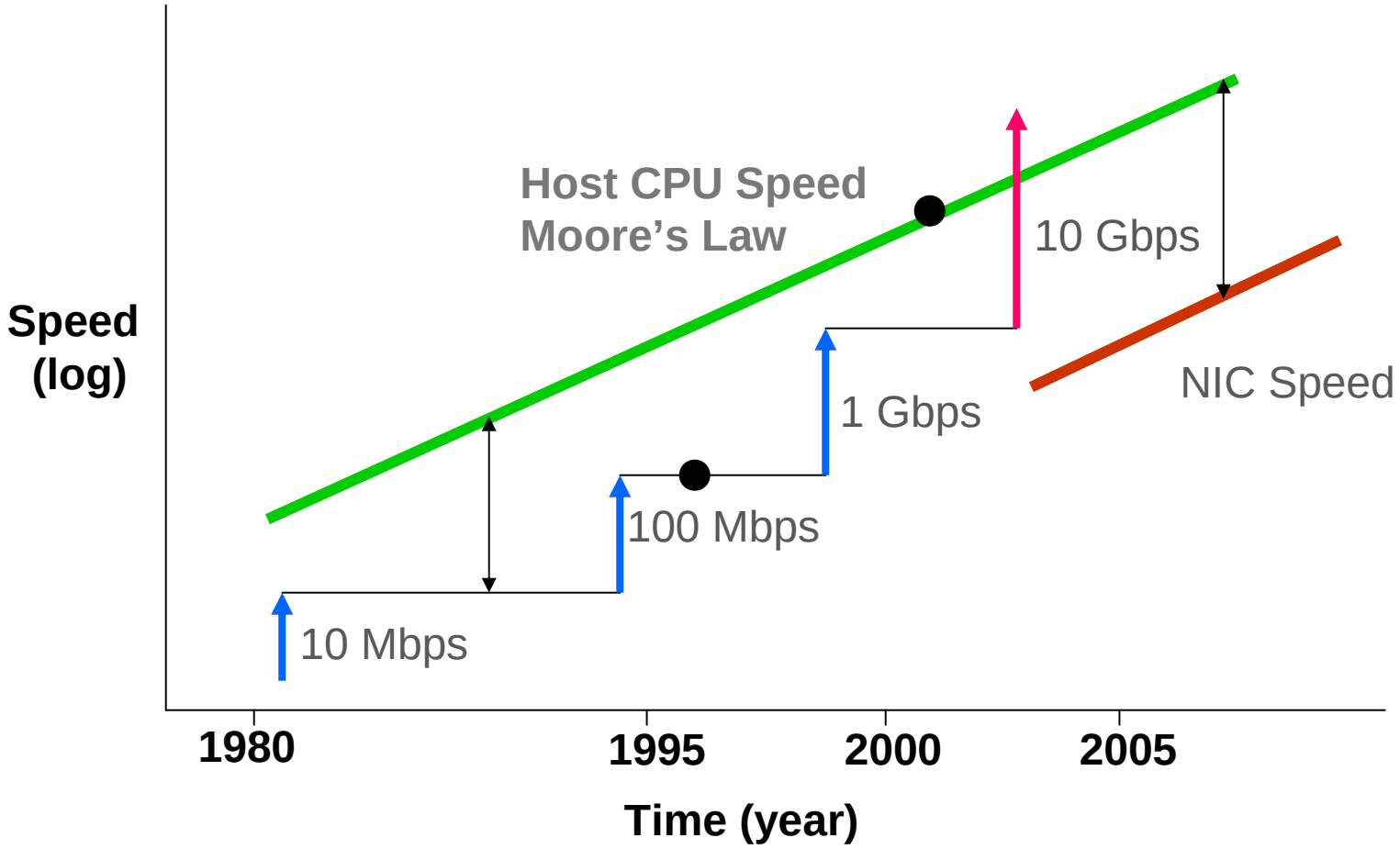
Offload controversy

- *Prasenjit Sarkar et al.,* “When does offload help?” (FAST 2003)
  - Host CPU fast enough
  - NIC becomes the bottleneck
- *Hennessy & Patterson, “*Pitfall: slow NICs”
- *Jeff Mogul*, “TCP offload is a dumb idea whose time has come” (HotOS 2003)
- *Offload NIC vendors*, “Any network application should look no further”

Tech trends

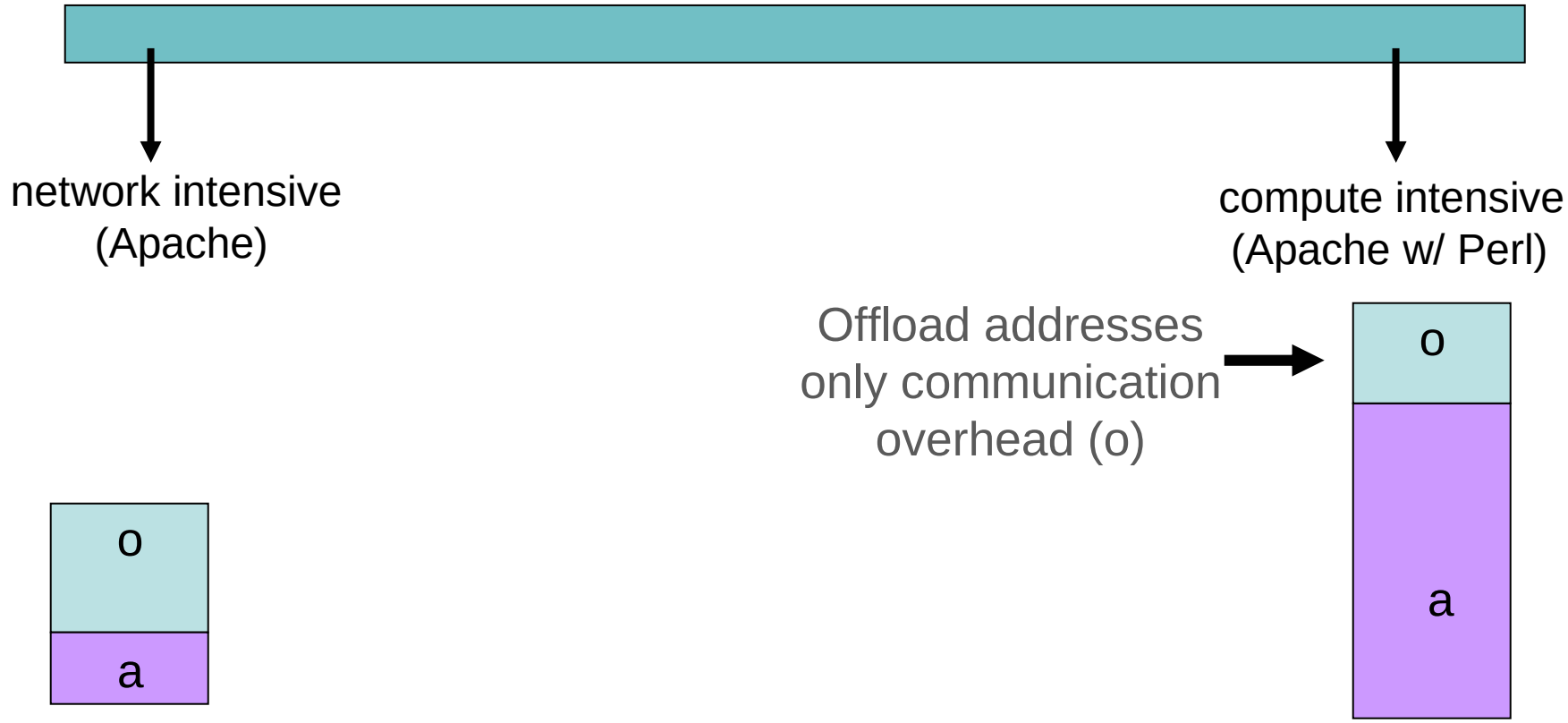

Host CPU Speed
Moore's Law
10 Gbps
NIC Speed
1 Gbps
100 Mbps
10 Mbps
Speed
(log)
1980
1995
2000
2005
Time (year)

Tech trends


Host CPU Speed
Moore's Law
10 Gbps
NIC Speed
1 Gbps
100 Mbps
10 Mbps
Speed
(log)
1980
1995
2000
2005
Time (year)

Application trends

Where do typical data center applications lie?



Motivation

- Continuum of tech trends and applications
  - Point studies can be misleading
  - Explore the whole space
  - Tech trends and applications change

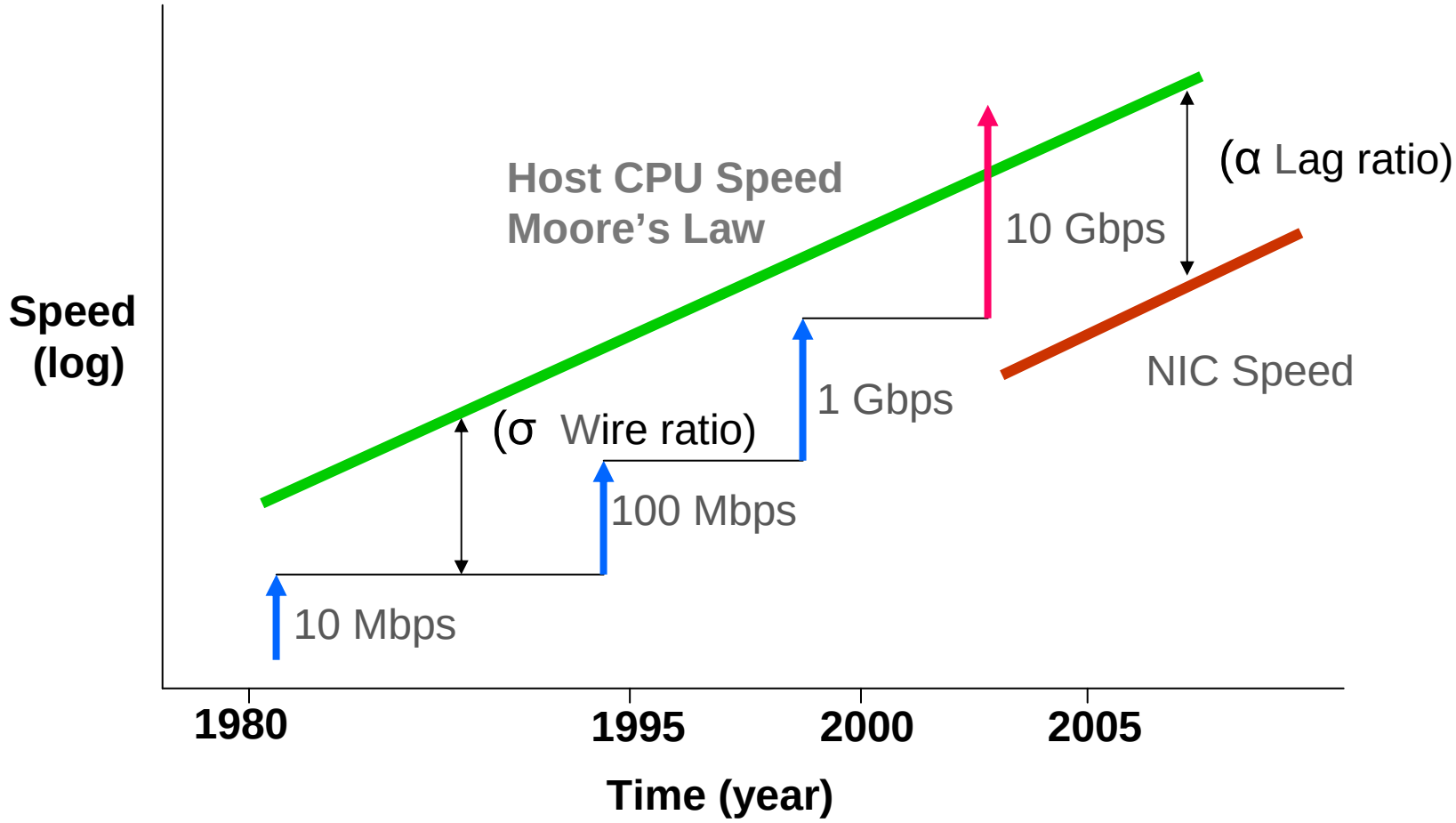
LAWS

- LAWS explores the whole space
  - Simple analytical model
  - Captures the entire continuum of applications and tech trends
  - Independent of any given point in tech and application space
  - Applicable to any low-overhead I/O technique (including non-IP SANs)

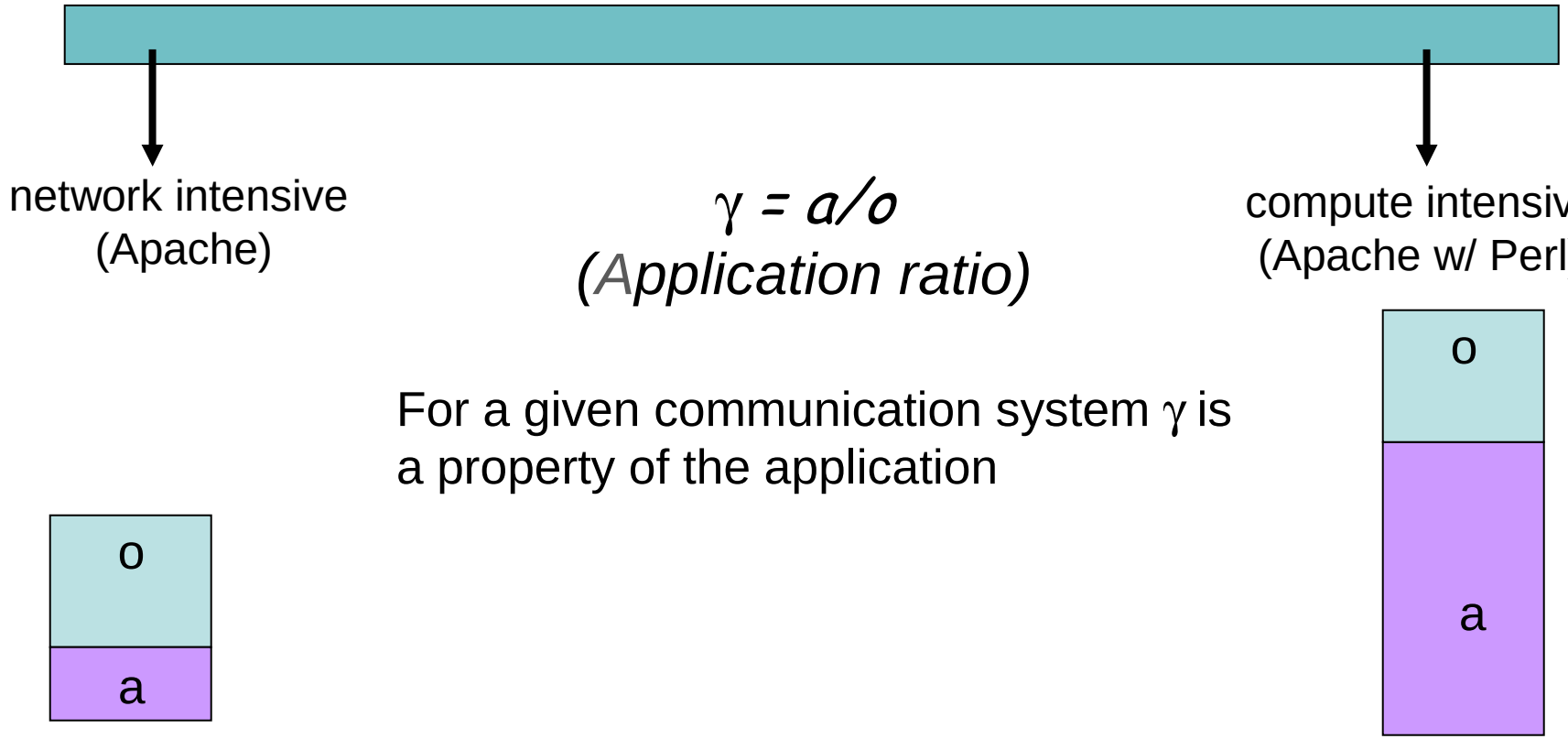
LAWS ratios

α	Lag	Ratio of Host CPU speed to NIC processing speed
γ	Application	CPU intensity of the application
σ	Wire	Percentage of wire speed the host can deliver for *raw* communication without offload
β	Structural	Portion of network work not eliminated by offload

LAWS captures tech trends



LAWS captures application trends

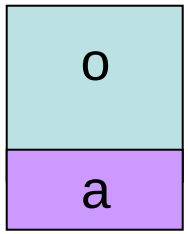


LAWS analysis

- Focus on throughput
  - Ignore latency
  - Internet servers are fully pipelined
- Metric
  - Throughput speedup (% increase in throughput) and *not throughput*
  - *(after – before) / before*

Throughput speedup (% increase in throughput)

- Very simple algebra
- (after – before) / before
- before = 1 / (a + o) (*host limited*)
- What happens after offload
  - Host limited
  - Network limited
  - NIC limited



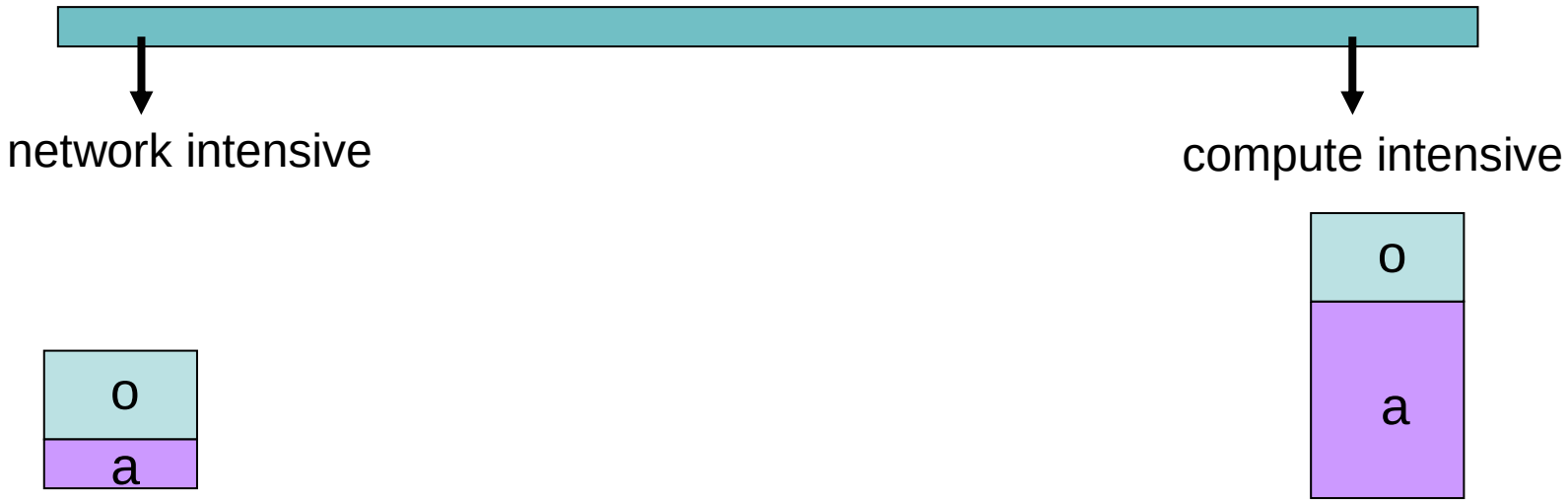
Host	1/a
Network	B
NIC	$1/o_{nic}$

B = network bandwidth

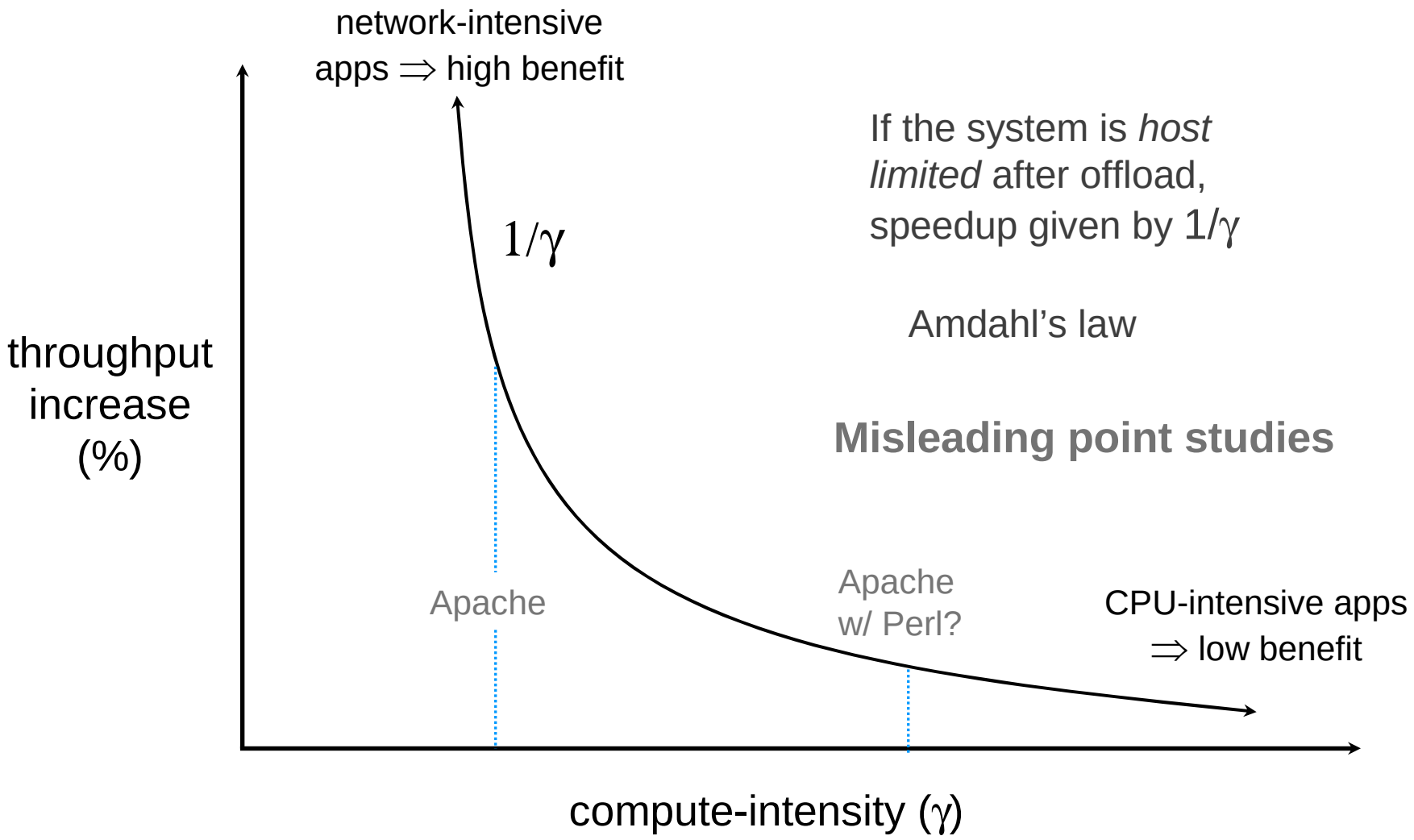
throughput = *min* (Host, Network, NIC)

Host limited

- Host saturated even after offload – host limited
- Captured by Application ratio (γ)
  - γ = a/o
    - higher γ means highly compute-intensive application



Host limited - benefits



If the system is *host limited* after offload, speedup given by 1/γ

Amdahl's law

Misleading point studies

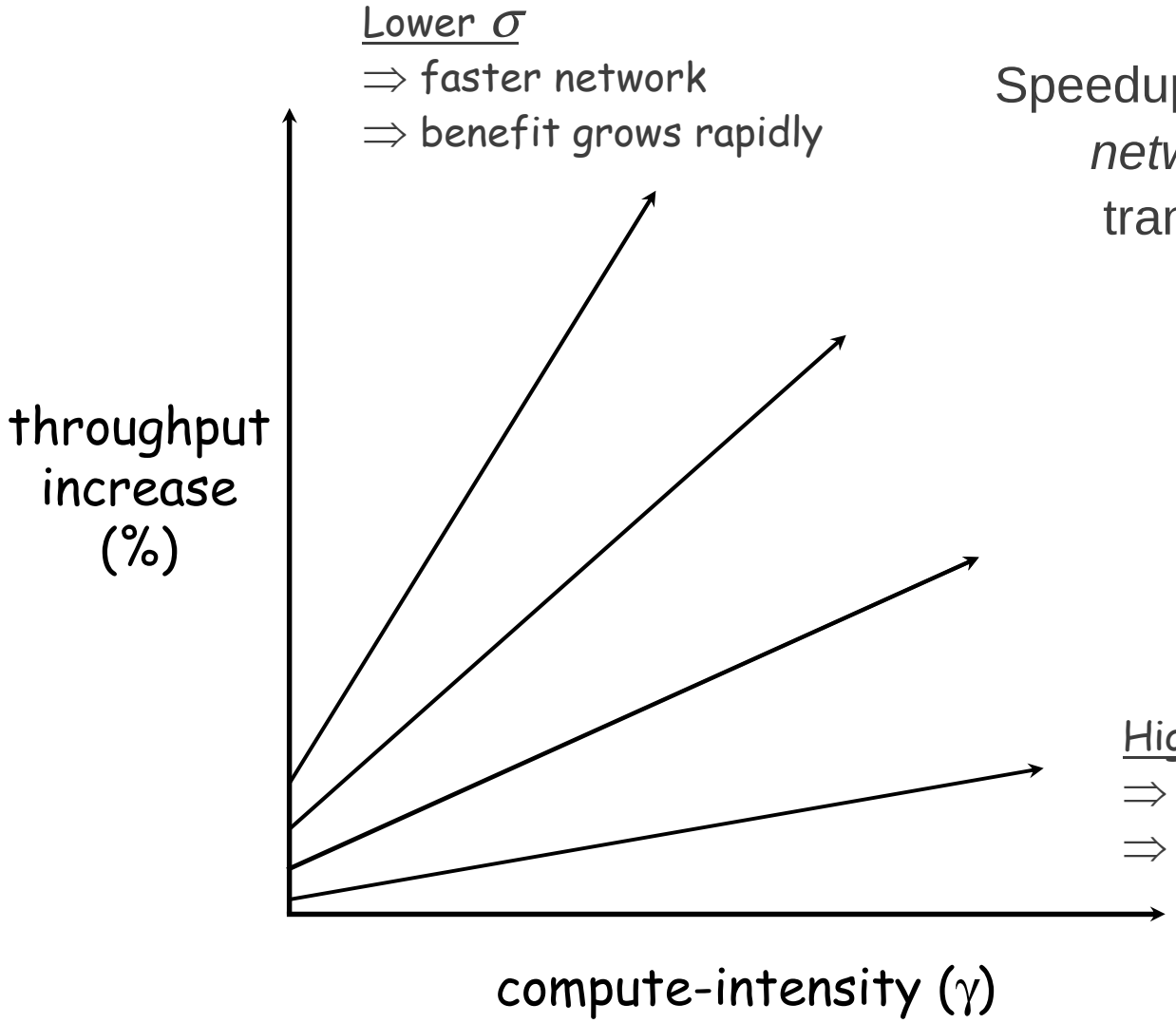
Network limited

- Network gets saturated after offload
- Captured by Wire ratio (σ)
  - σ represents the percentage of wire speed delivered by host for raw communication
  - o = communication processing overhead at host per unit of bandwidth
  - 1/o = the host throughput for raw communication
  - B = network bandwidth
  - $\sigma = (1/o) / B$

Understanding σ

- σ captures how slow/fast the network is compared to host
- Low σ - fast network ($\sigma < 1$)
- High σ - slow network ($\sigma >> 1$)
- $\sigma = 1$ - host matched to the network, and hence ‘realistic’

Network limited - benefits

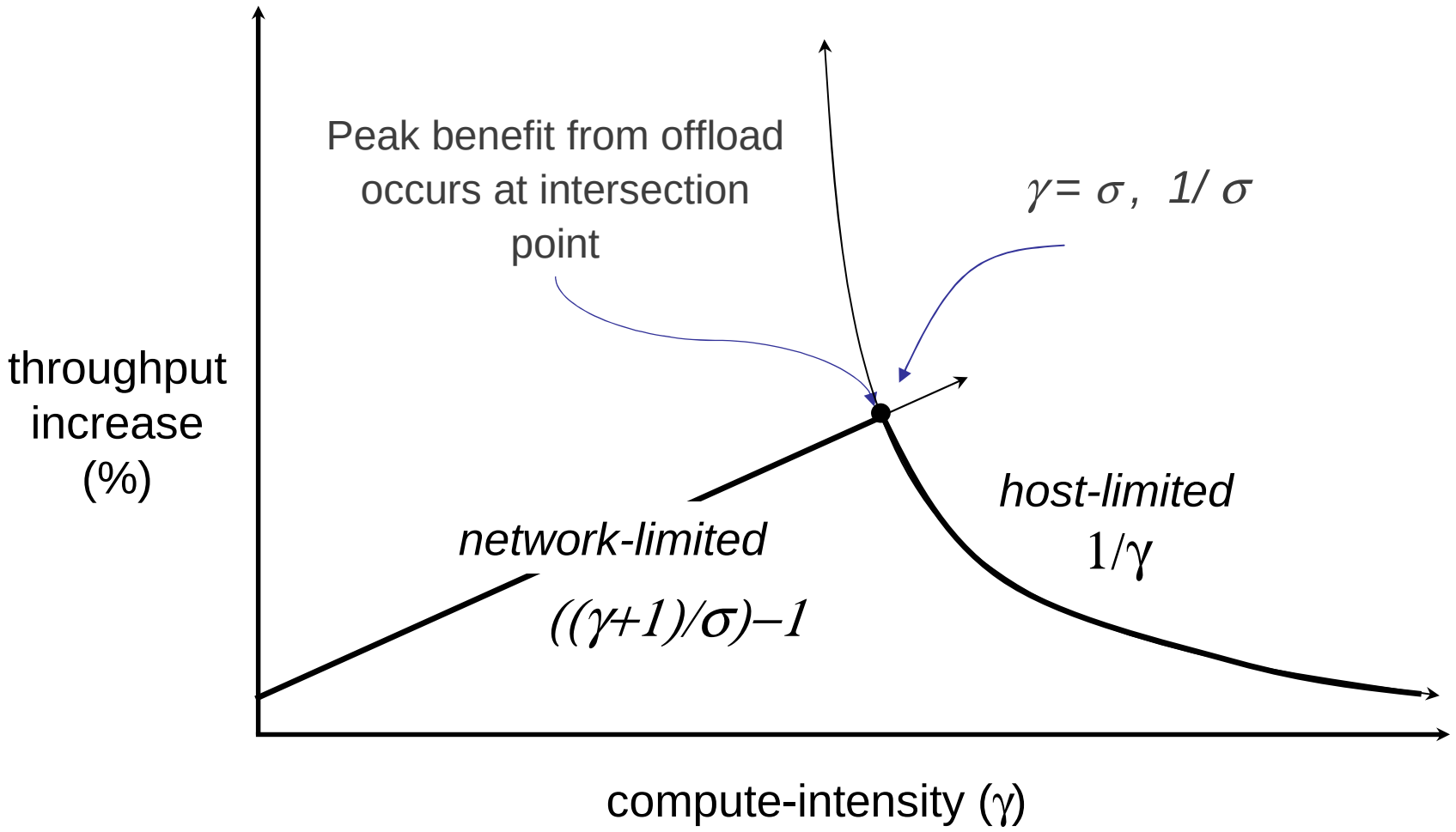


Speedup when the system is *network-limited* after transport offload is:

$$((\gamma+1)/\sigma)-1$$

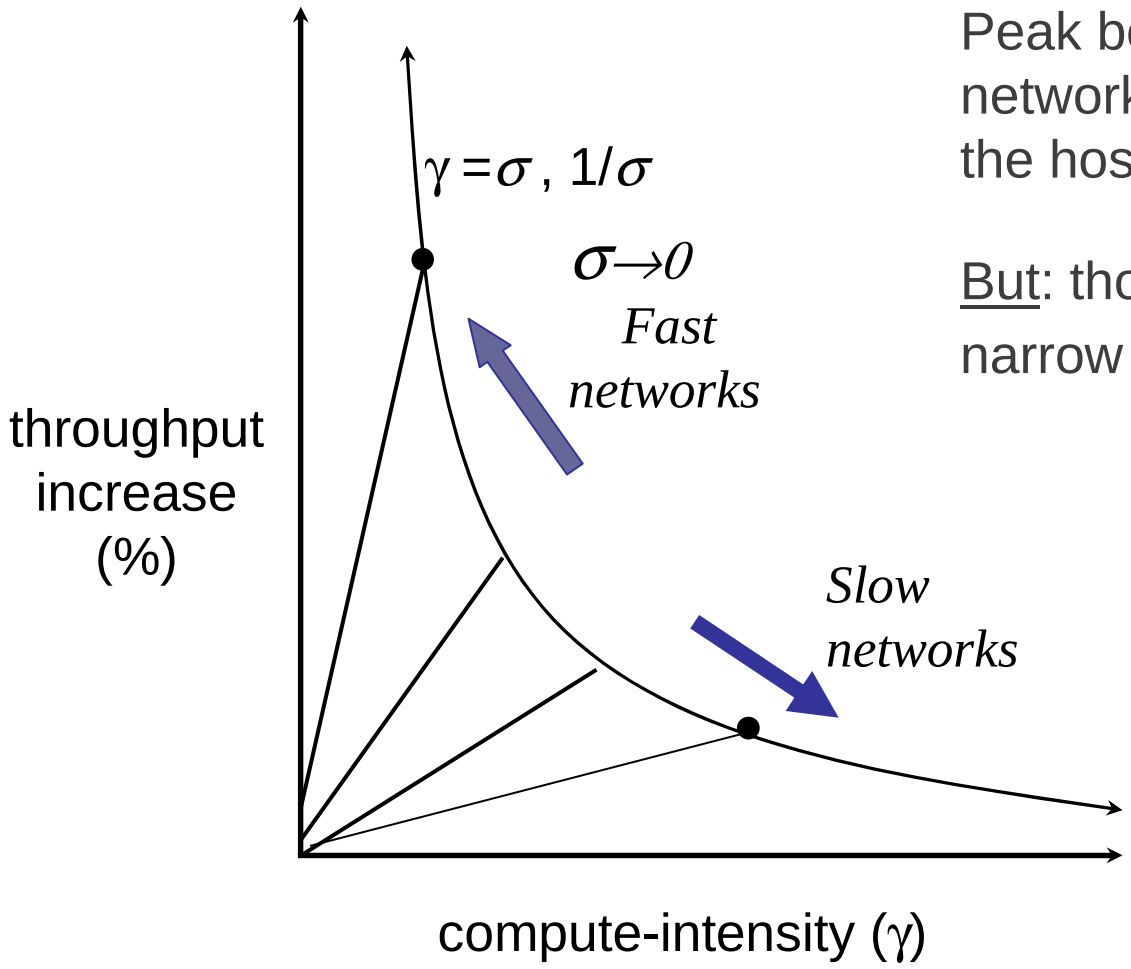
$$\sigma = (1/o) / B$$

Putting it all together



Offload for fast/slow networks

$\sigma = (1/o) / B$

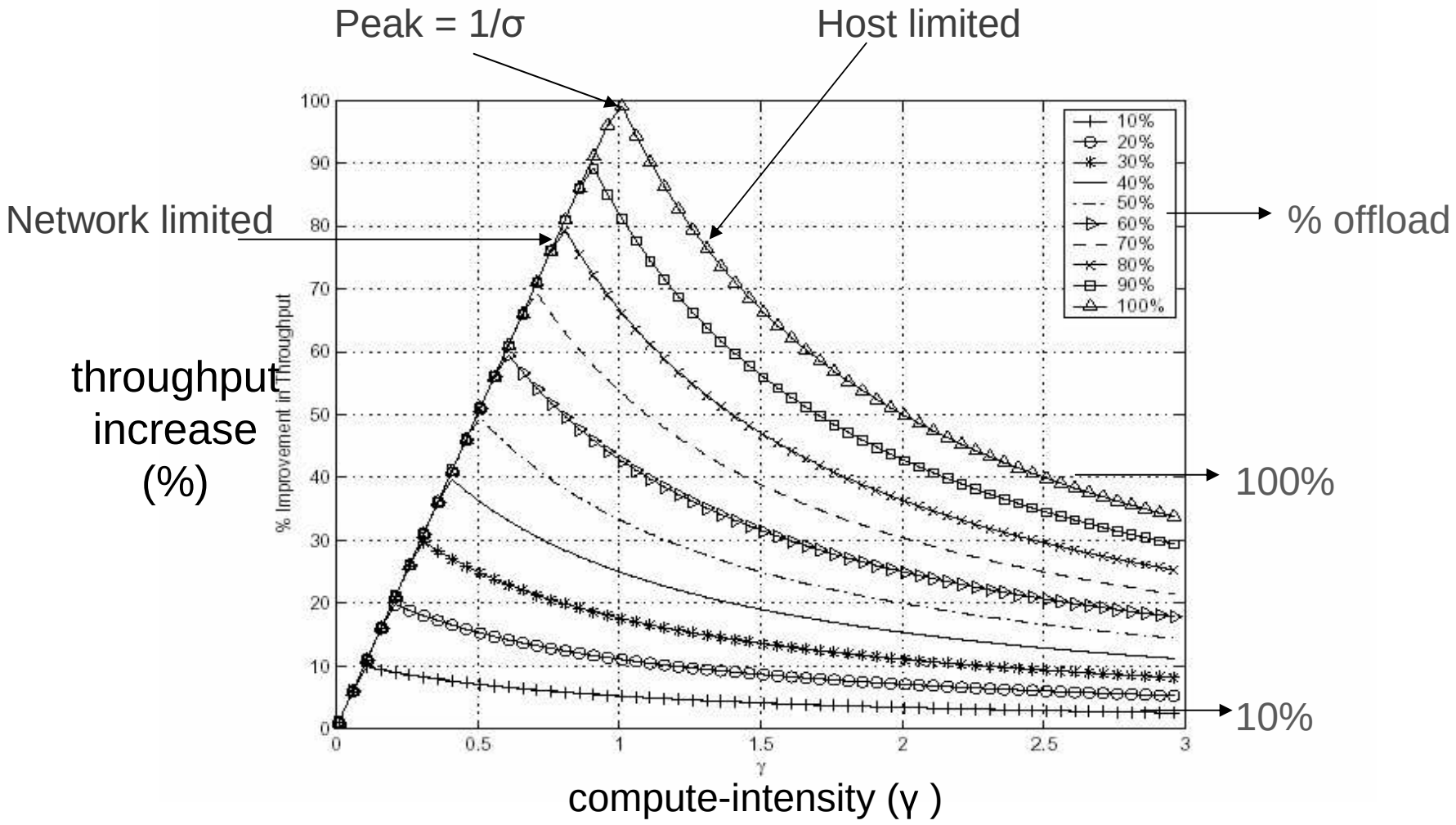


Peak benefit is unbounded as the network speed advances relative to the host!

But: those benefits apply only to a narrow range of low-γ applications.

Will network bandwidth outrun Moore's law?

Offload for ‘realistic’ network (σ = 1)

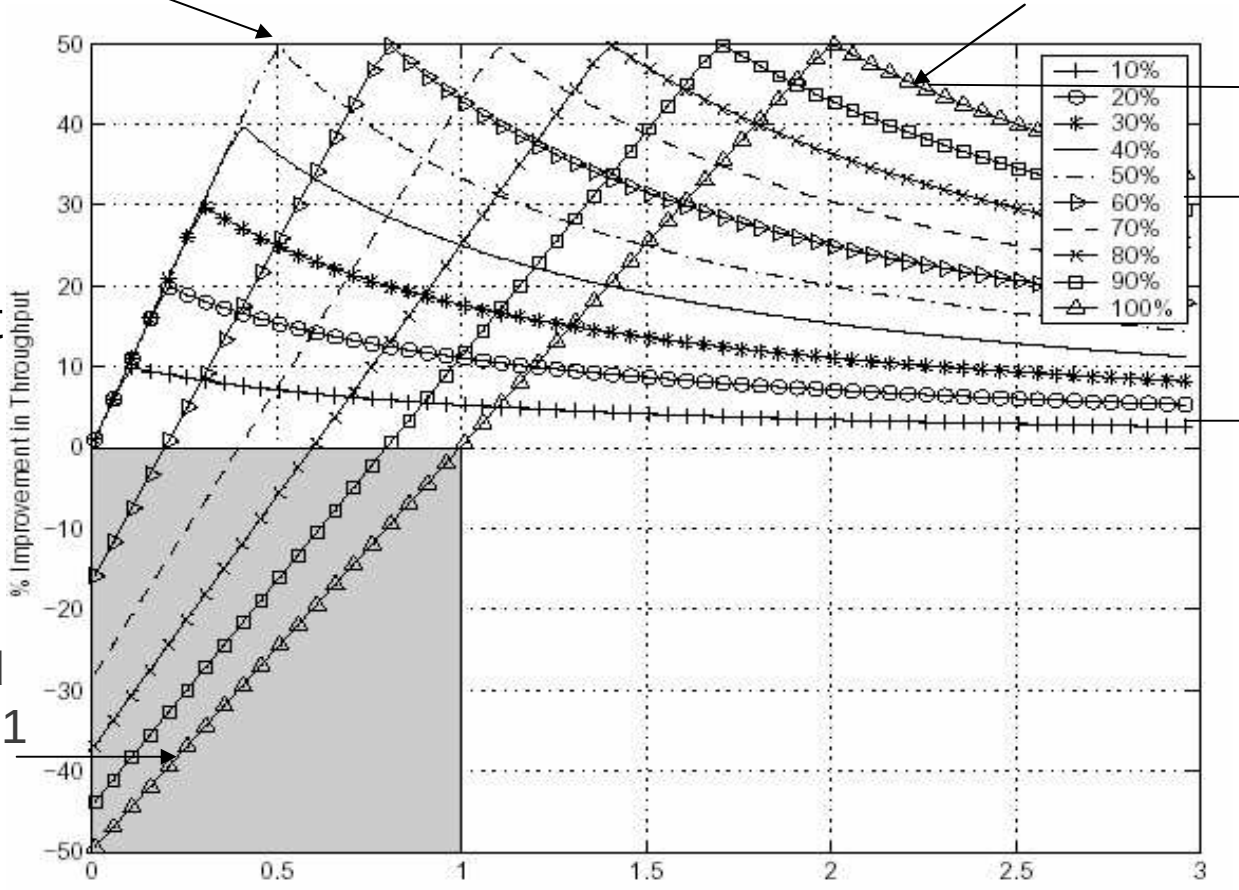


- **Peak benefit bounded by a factor of 2 in realistic case**
- **More offload does not help when *not* host limited**

NIC limited

- So far assume only Network or Host can be the bottleneck
- What happens when NIC gets saturated?
- The *lag ratio* (α) captures the *relative* speed of the host and NIC for communication processing
  - α = 2 means that NIC is half as fast as host CPU

Benefits - slow NIC, α = 2



Pitfall: Slow NIC (Hennessy & Patterson)

- **Peak benefit reduces by half**
- **Some applications may be harmed by offload**

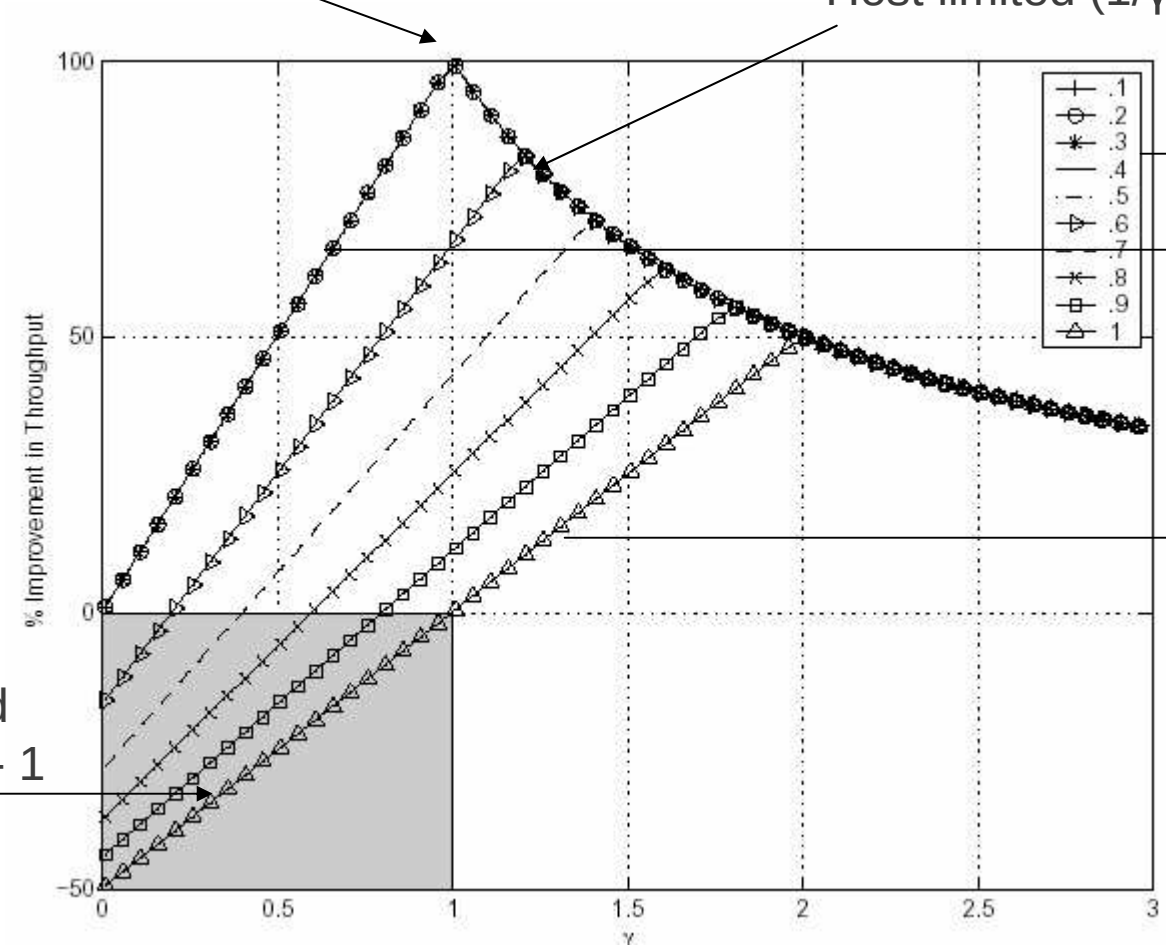
RDMA / DDP

- So far assumed exact offload
  - work remains the same after offload
  - RDMA (Direct Data Placement) allows structural change which cuts down the work
  - Captured by β
    - amount of *work remaining* on the NIC after offload

Example

- $\alpha = 2$
  - NIC is half as fast as host
  - 18 months behind the host CPU (Moore's law curve)
- $\beta = 0.5$
  - Only 50% work remains!
- Equivalent to a NIC that is twice as fast as the current NIC
  - Gained 18 months on the Moore's law
  - Eased time to market pressure for offload NICs

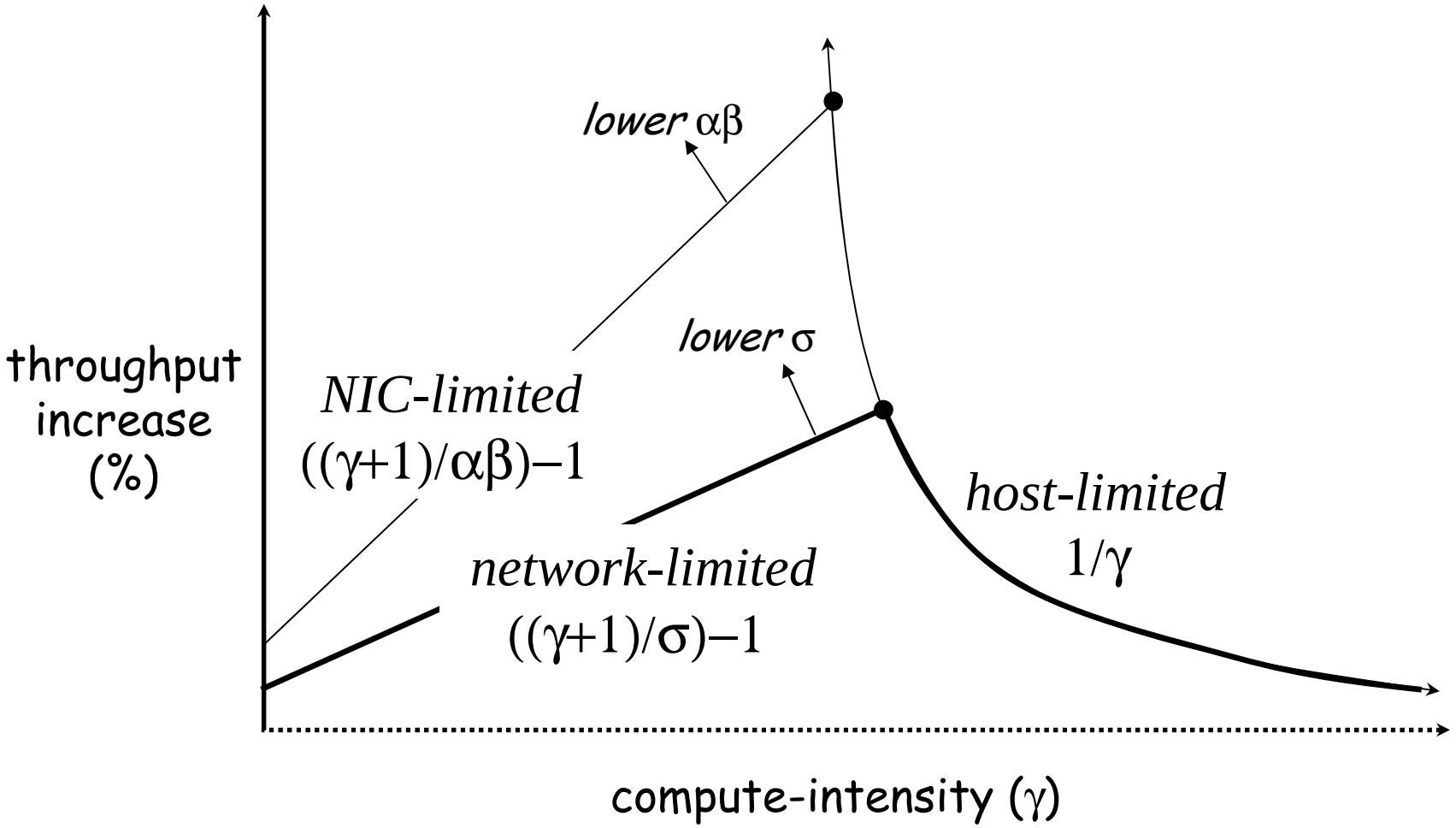
Benefits – slow NIC with DDP



Jeff Mogul, "TCP offload is a dumb idea whose time has come"

- **Peak benefits increases to 100% for same slow NIC**

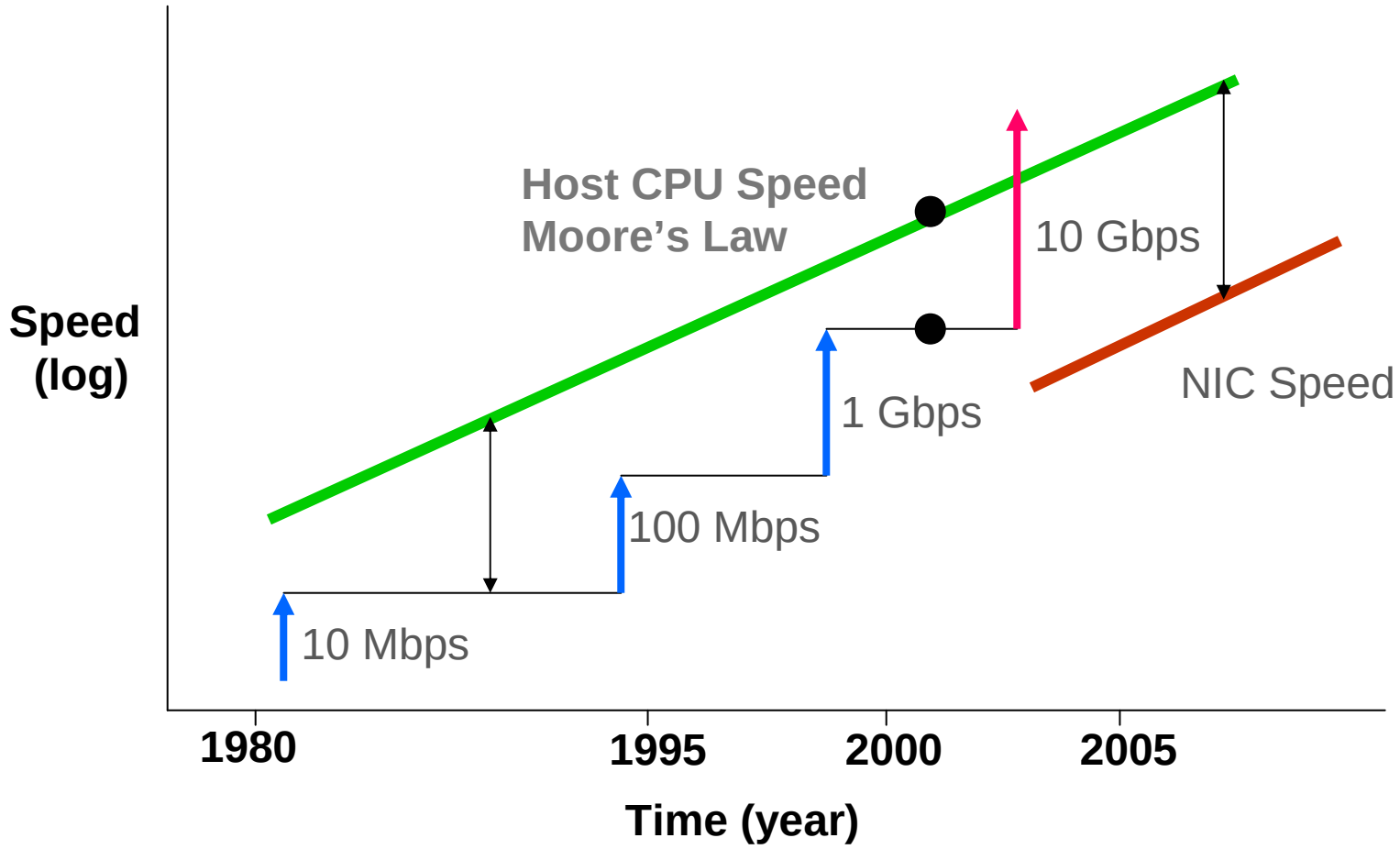
Overall Picture



Conclusion

- To understand the role of TCP/IP offload, RDMA, etc. we need to understand the applications (γ)
- Point studies are misleading
  - Choosing γ and σ carefully can yield very high benefits
  - But those benefits will be elusive in practice
- LAWS analysis exposes fundamental opportunities and limitations of offload and other approaches to low-overhead I/O (including non-IP SANs)
- Helps guide development, evaluation, and deployment

Tech trends


Host CPU Speed
Moore's Law
10 Gbps
NIC Speed
1 Gbps
100 Mbps
10 Mbps
Speed
(log)
1980
1995
2000
2005
Time (year)

Example

- σ = *(1/o) / B*
- Gigabit ready host attached to 10GigE
  - σ = .1 – fast network
- Gigabit ready host attached to 100Mbps
  - σ = 10 – slow network
- Gigabit ready host attached to Gigabit network
  - σ = 1 – ‘realistic case’

Capturing host/network gap

- Captured by Wire ratio (σ)
  - σ brings out the percentage of wire speed delivered by host for raw communication
  - o = host communication overhead per unit of bandwidth
  - 1/o = the host throughput for raw communication
  - B = network bandwidth
  - $\sigma = (1/o) / B$

Capturing host/network gap