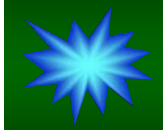
Dragon Slayer Consulting

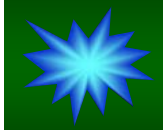
Introduction to the Value Proposition of InfiniBand

Marc Staimer – marcstaimer@earthlink.net
(503) 579-3763



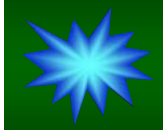
Introduction to InfiniBand (IB) Agenda

- IB defined
- IB vs. FC & GbE
- IB architecture
- Real market problems IB solves
- Market projections
- Conclusions



Definition of Input/Output

- "The transfer of data into and out of a computer"
  - *Maintain data integrity*
  - *Protect all other data in the computer from corruption*
  - *Through the use of Operating System defined mechanisms*
    - Usually



Three (3) Distinct Classes of I/O

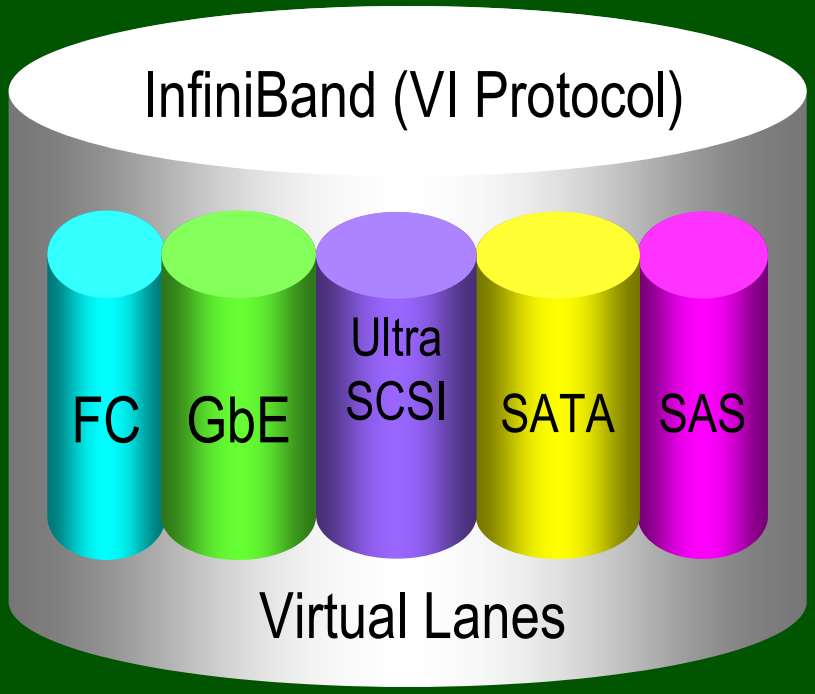
- Block protocol
  - *Typically disk oriented*
- Network protocol
  - *Typically IP oriented*
- Inter-Process Communication
  - *IPC*

Characteristics I/O Classes

	Block Protocol	Network Protocol	IPC
Latency Tolerance	Dozens of milliseconds	100s of Milliseconds	Dozens of Microseconds
Avg Message Size	Very large	Small to large	Small to large
Context	Data center/campus FC	Global	Server cluster/data center
Predominate Protocol	Fibre Channel Protocol (FCP)	Ethernet / TCP/IP	Emerging - VI

IB Defined

- The 1st unified, simplified, & consolidated I/O Fabric
  - *Designed from the ground up for all aspects of I/O*
  - *Shared memory vs. shared bus*
  - *Leverages virtual lanes or pipes*
    - ➥(multiple fabrics in one)
  - *Spec'd for today & tomorrow*
    - ➥1x = 2.5Gbps
    - ➥4x = 10Gbps
    - ➥12x = 30Gbps
  - *Native VI protocol*
    - ➥OS bypass
  - *Credit based flow control*
  - *Key: extends server I/O*
    - ➥Outside the box

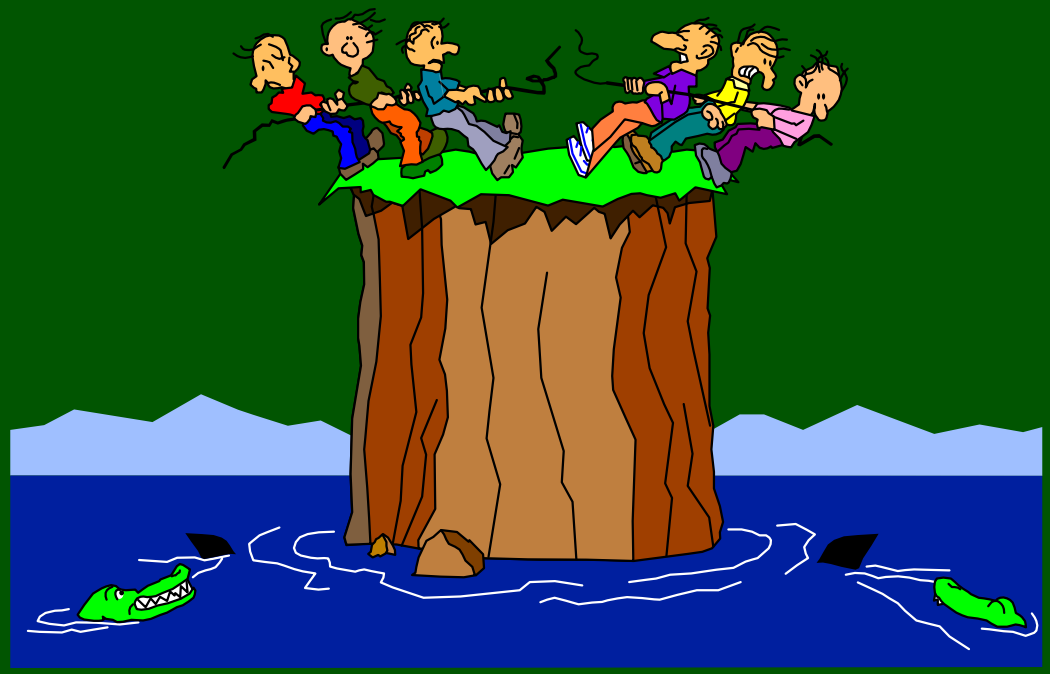


Why Do We Need Yet Another Fabric?

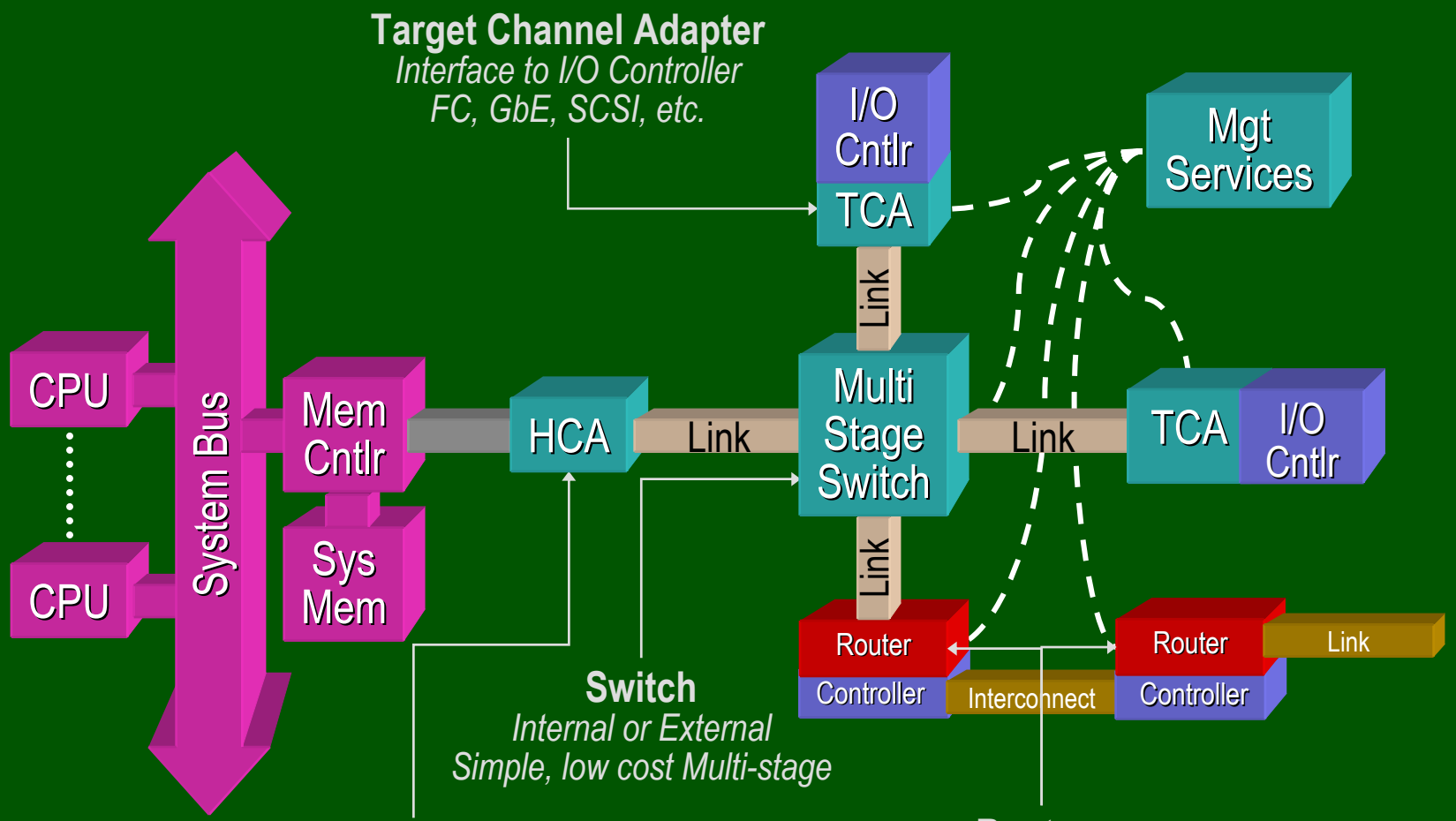
- The issue is not the fabric, the issue is server I/O
- Current GbE & FC fabrics do not solve server I/O bottlenecks
  - *Bus contention*
- GbE & FC fabrics weren't specifically designed for clustering
  - *They can do it…AND*
    - ➥Message queue depths and performance not optimal
    - ➥Performance is often inadequate

IB vs. FC vs. GbE Conclusion

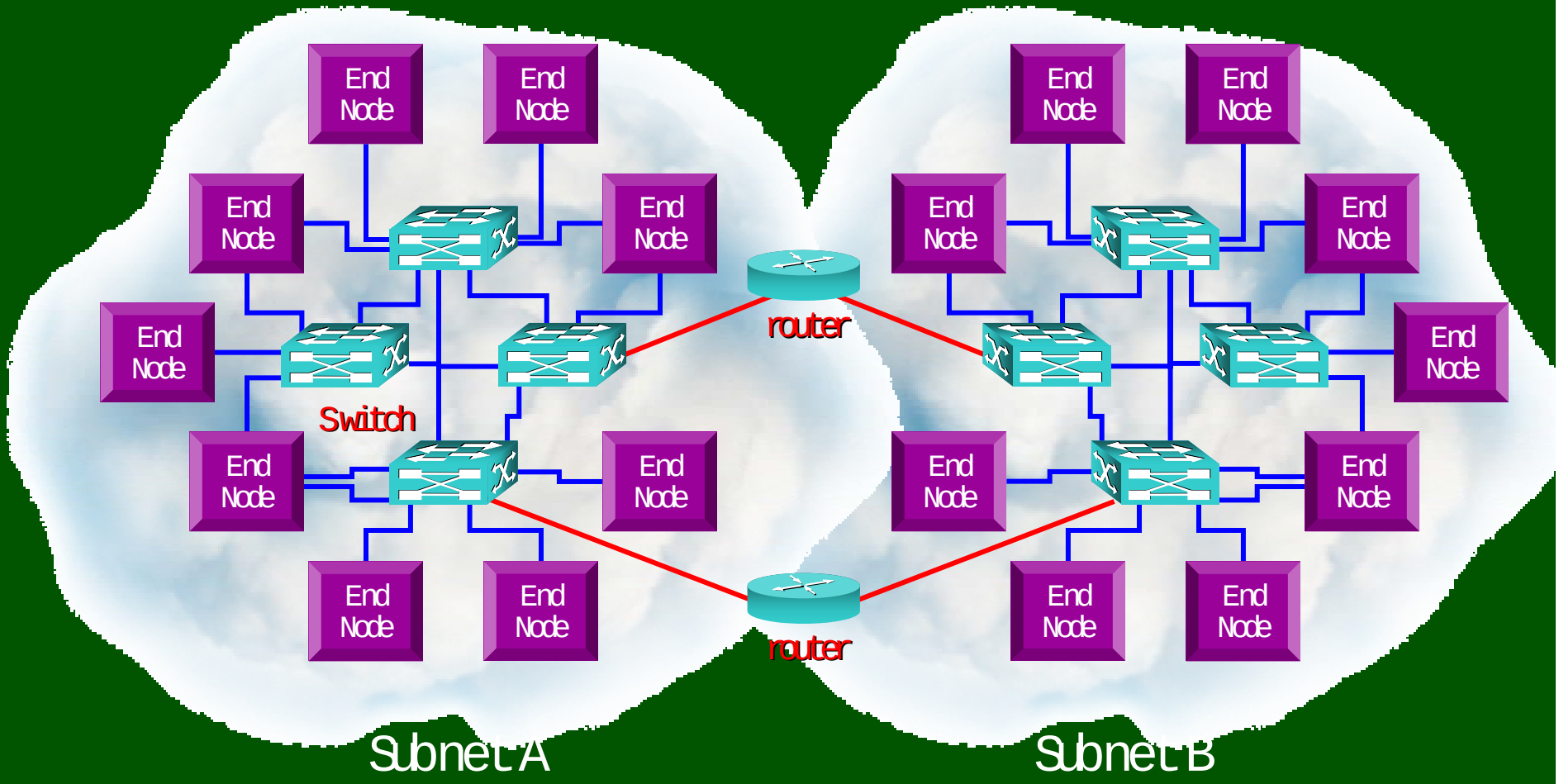
- Initially complimentary – IB will not replace FC or GbE
  - *Investment protection*
- Eventually competitive and complimentary
  - *They will compete for some of the same budget dollars*



IB Architecture

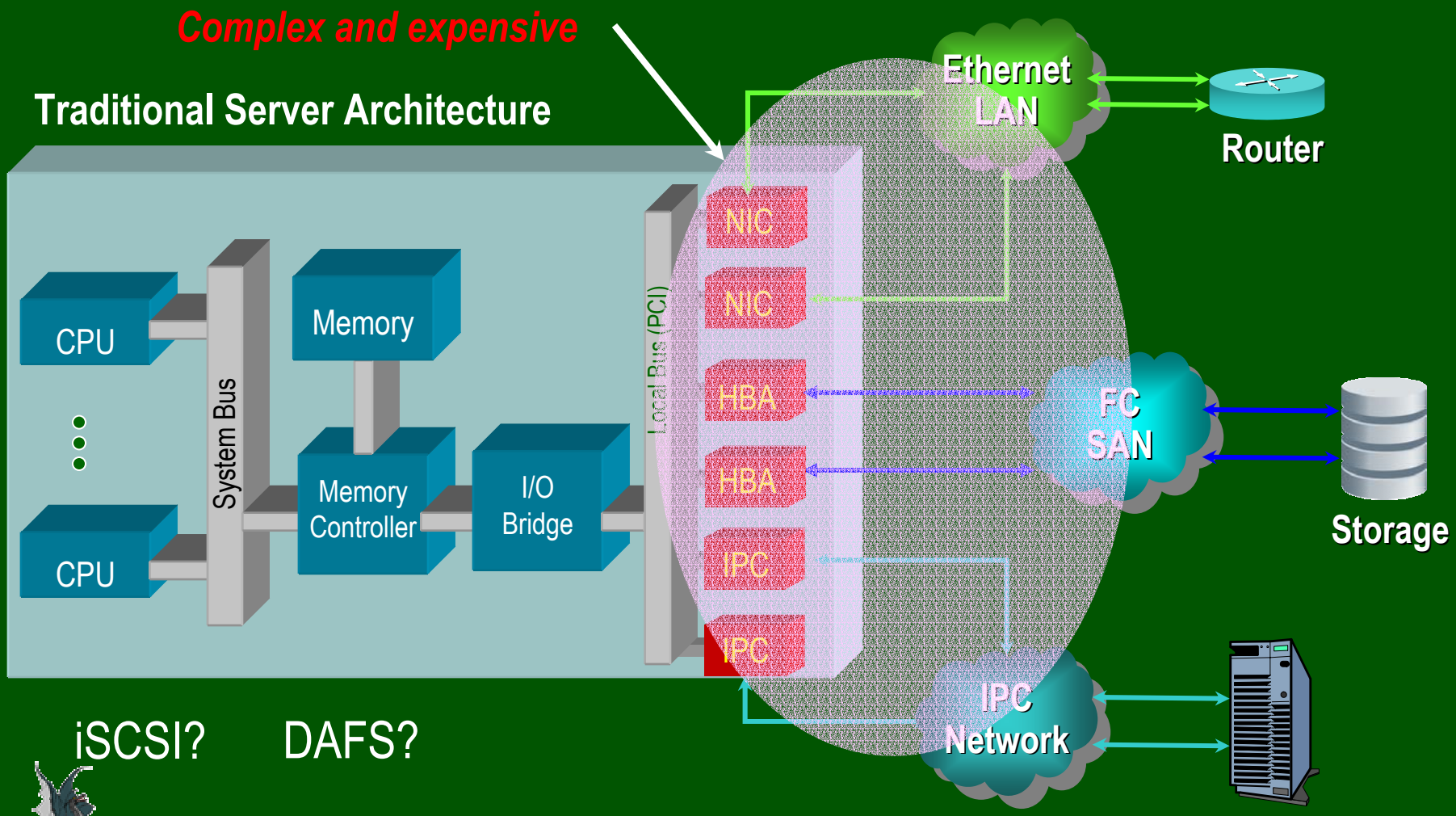


IB Fabric BW Increases as Switches are Added



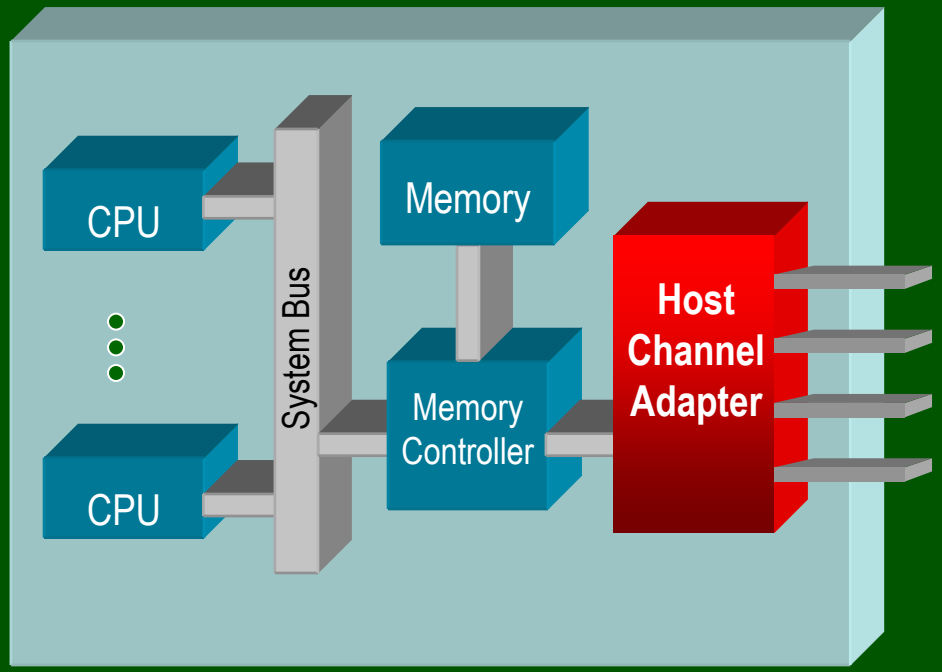
I/O Architecture Today

- Traditional Server & Infrastructure w/dedicated I/O



InfiniBand Based I/O

InfiniBand Server Hardware Architecture

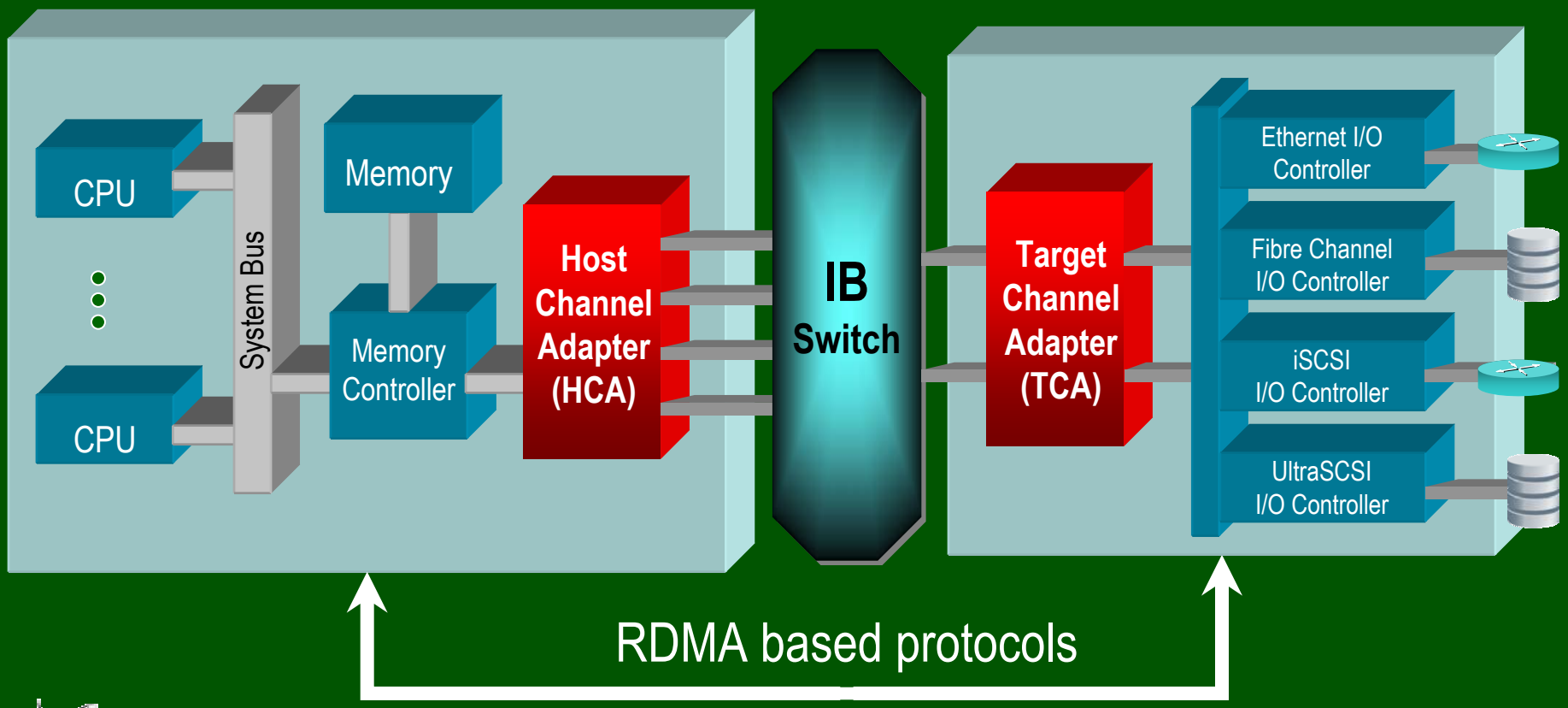


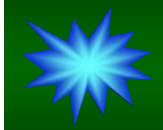
Multiple IBA links

- 2.5 Gbps
- 10 Gbps

Solve redundancy problem once

InfiniBand Based I/O

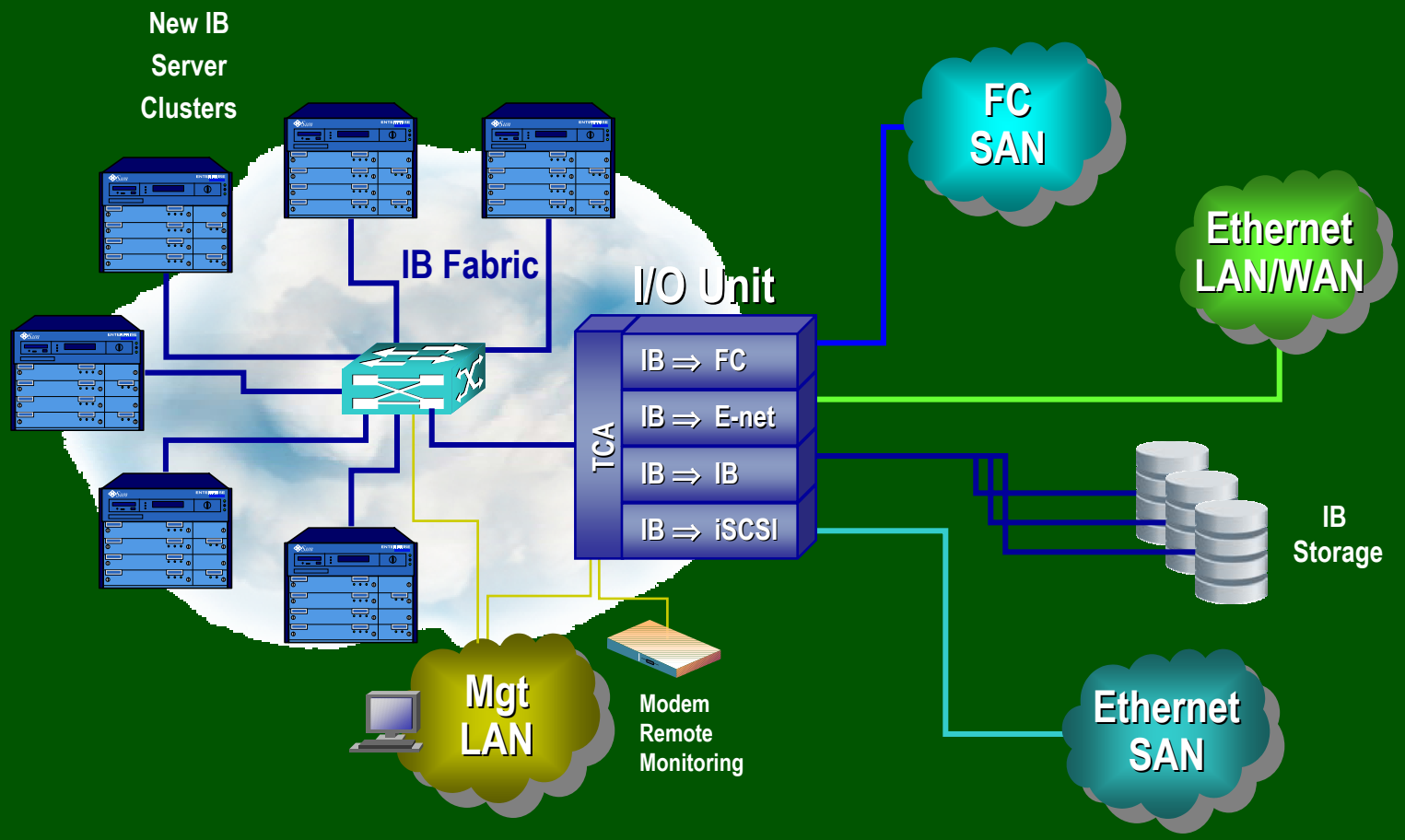




Market Problems IB Solves

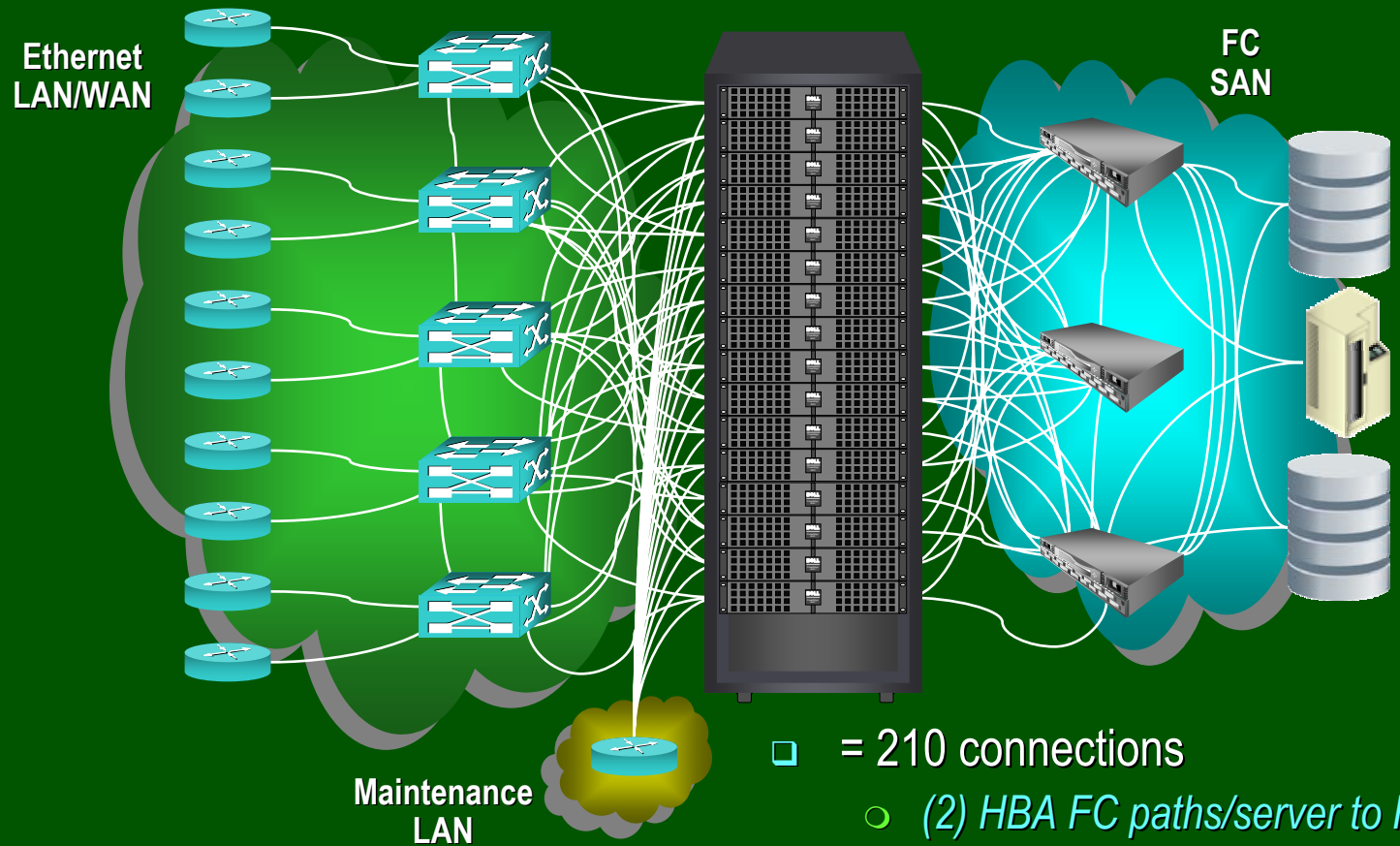
- Higher performance lower cost I/O (Shared I/O)
  - *Converges clustering, networking, & storage into one fabric*
    - The IAN (I/O Area Fabric)
  - *Reduces:*
    - IT management tasks
    - Server workloads
    - TCO
- PCI Bus I/O constraints
- Low cost HP/HA server clustering
  - *Lowers the cost of server blade systems*
    - Enables higher density server blade clusters

Higher Performance Lower Cost I/O (Shared I/O)



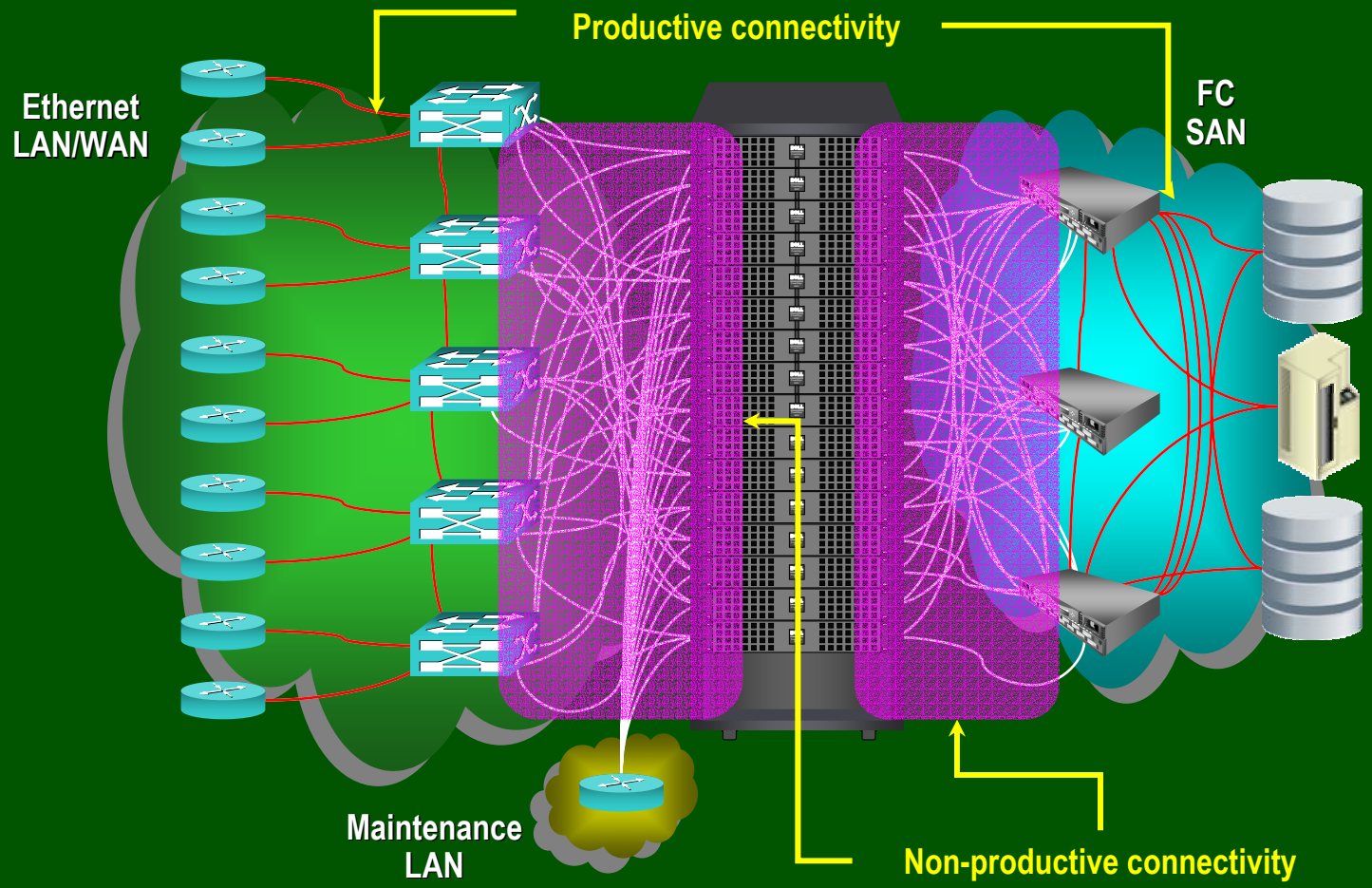
Current High Availability I/O Configuration

- 16 Rack mount servers with dedicated I/O per server



- = 210 connections
  - *(2) HBA FC paths/server to FC fabric*
  - *(4) FC paths to storage to FC fabric*
  - *(2) Ethernet paths/server to network*
  - *(1) Ethernet maint path/server to network*

Non-Productive Costly Connectivity

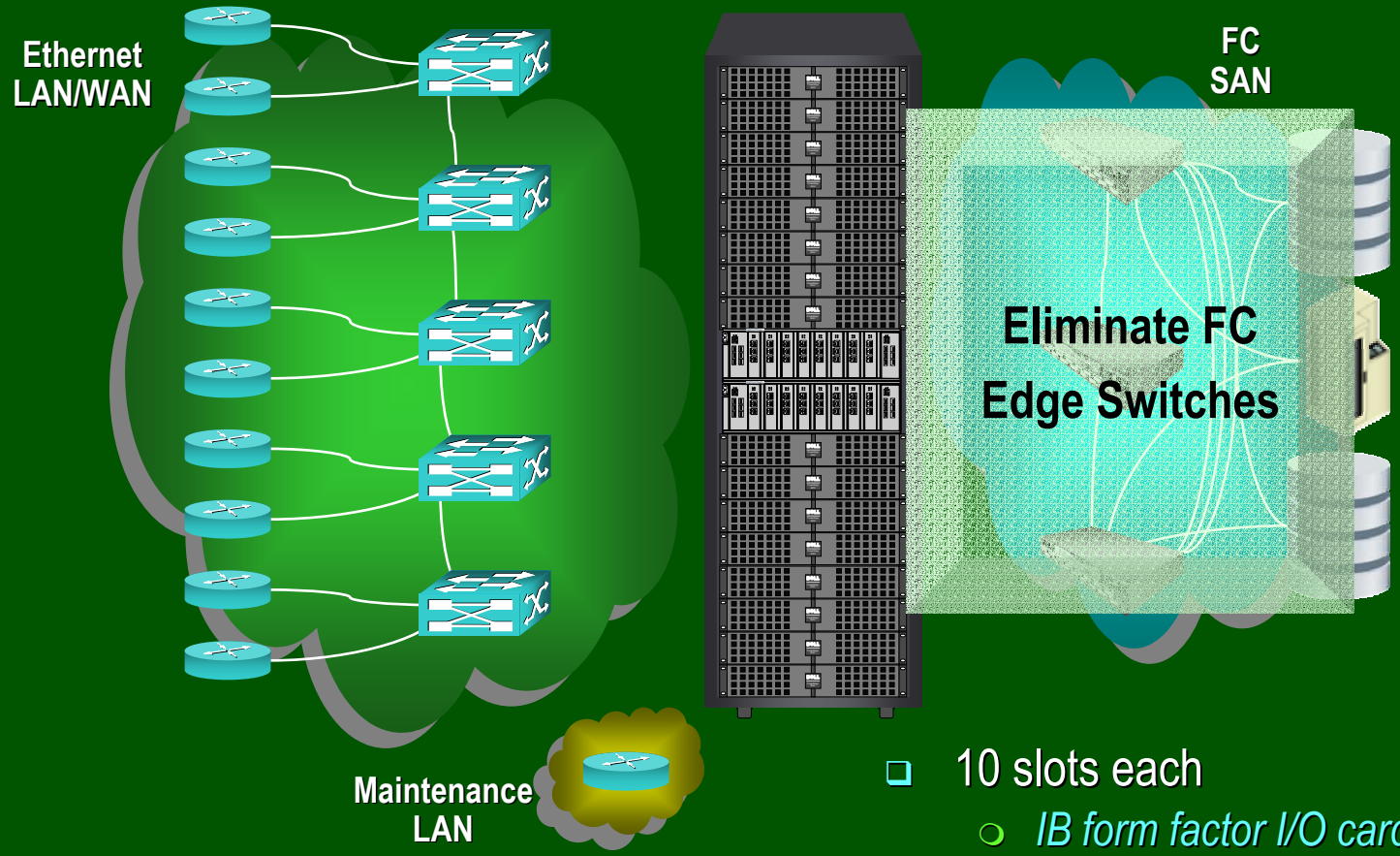


InfiniBand Shared I/O Chassis Example

- 19” rack mount environment
- 3U high
- IBA single high single wide std
- Integrated IBA fabric
- Up to 45 watts / linecard slot
- Hot swappable components
- Chassis Management Entity (CME)

IB Enabled High Availability Shared I/O

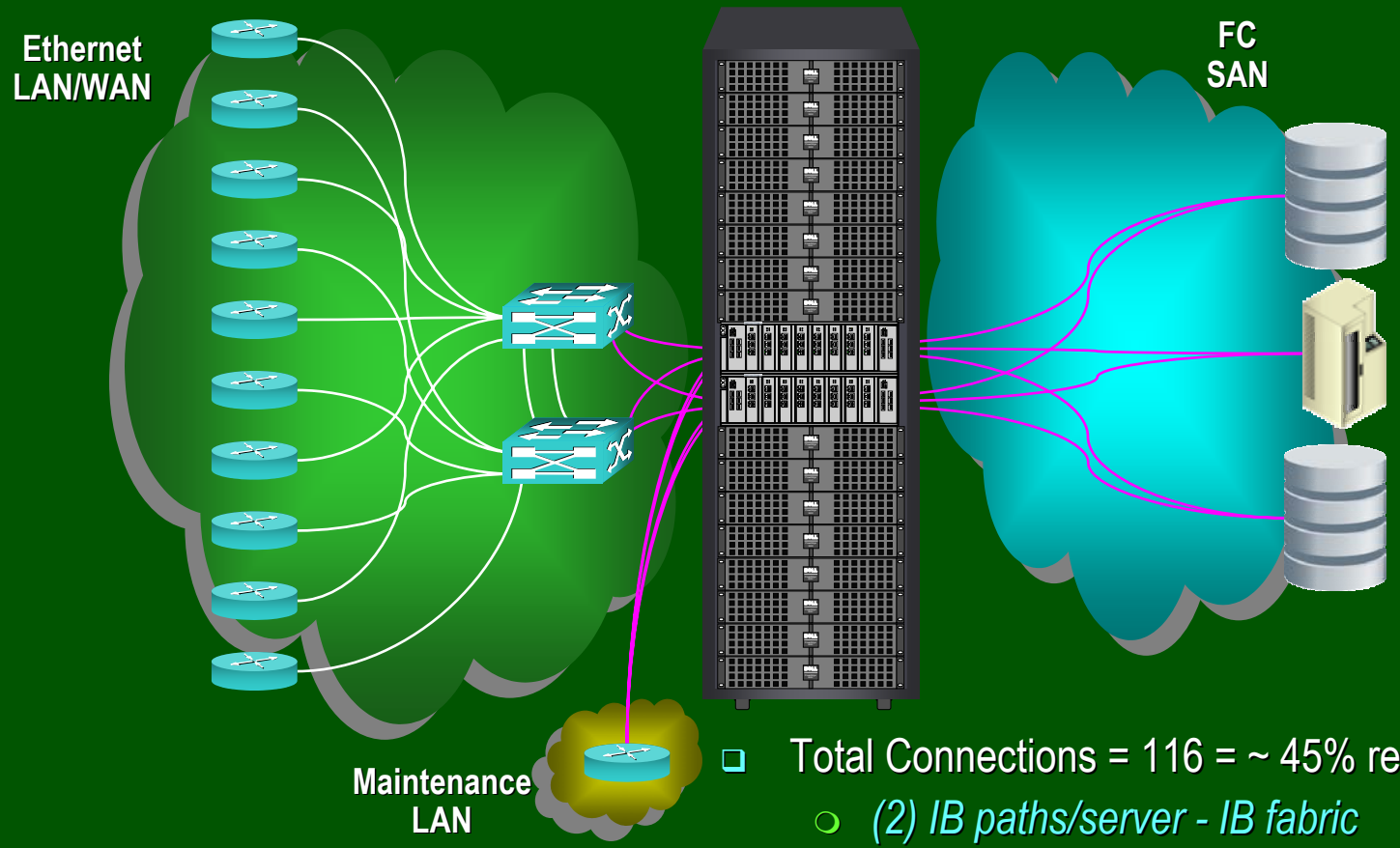
- Add dual redundant IB I/O Chassis



- 10 slots each
  - *IB form factor I/O cards*
  - *Multi-protocol*
    - ➥ FC, GigE, FastE, iSCSI, etc.
  - *Eliminate FC edge switches*

IB Enabled High Availability Shared I/O

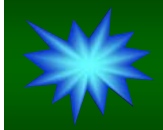
- Reduces LAN switch requirements



- Total Connections = 116 = ~ 45% reduction
  - *(2) IB paths/server - IB fabric*
  - *(6) FC paths to storage - FC fabric*
  - *(2) Ethernet paths/I/O subsystem – network*
  - *(2) E-net maint path/I/O subsystem - network*

Potential Savings

- Current dedicated I/O subsystem/server
  - *Costs = ~ $225,000*
- IB shared I/O System with
  - *Improved*
    - ➥BW, connectivity, manageability, availability
  - *Costs = ~ $112,500*
  - *Savings = ~ 50%*
- Additional non-hardware TCO gains
  - *Operational Expense*
    - ➥Estimated at 3x – 8x Capital Expense reduction
  - *Simpler system design to manage*

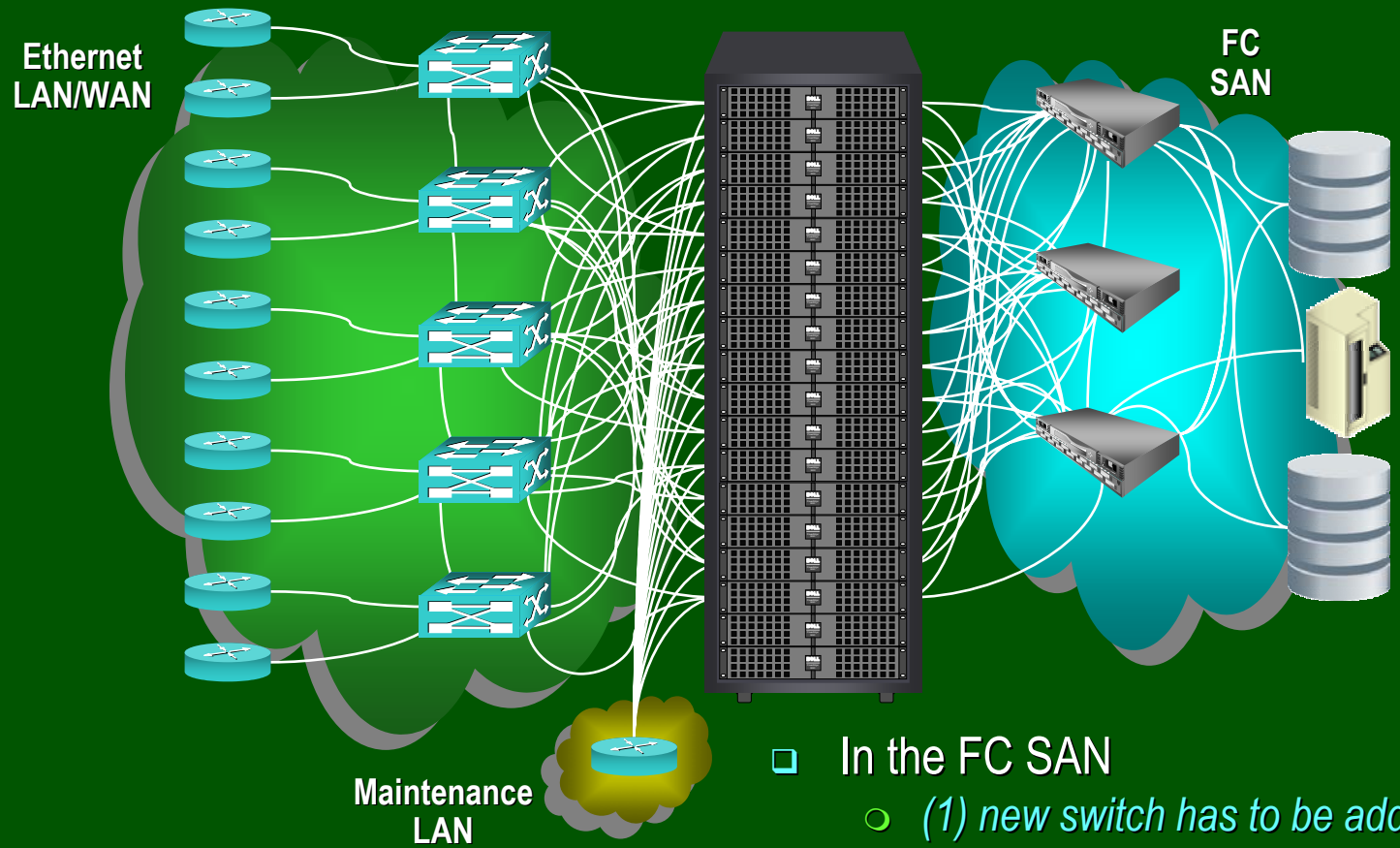


System Benefits

- Increased BW & connectivity per server
- Reduced infrastructure complexity
- Reduced power & space
- BW migration to bursting servers
- Natural low latency IPC network

Managing Scalability w/Traditional I/O

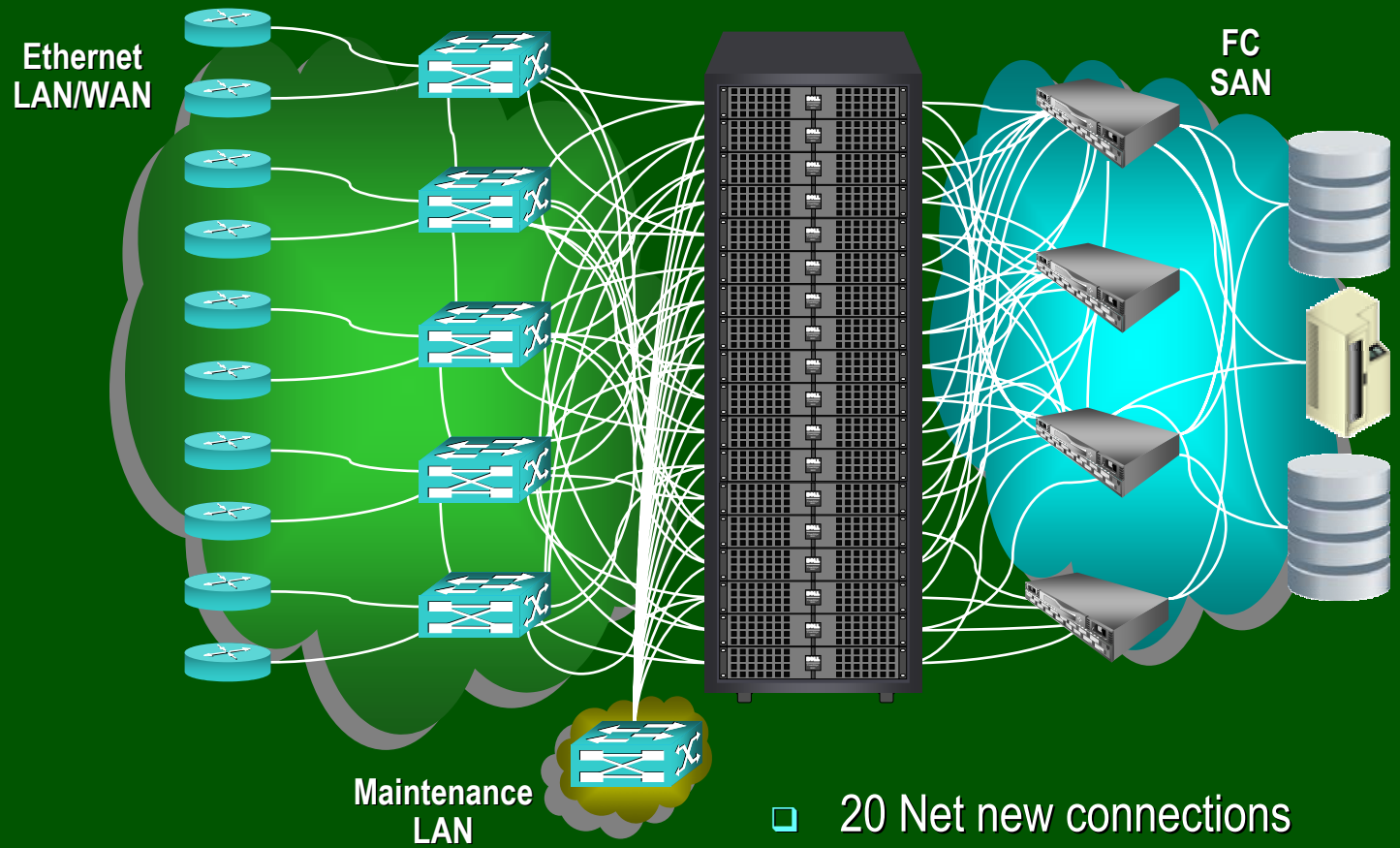
- What happens when just 2 more servers are added?



- In the FC SAN
  - *(1) new switch has to be added*
  - *Fabric will need to be reconfigured*
- Maintenance LAN will also need to change
  - *From a 16-pt switch/router to 24-port*

Managing Scalability w/Traditional I/O

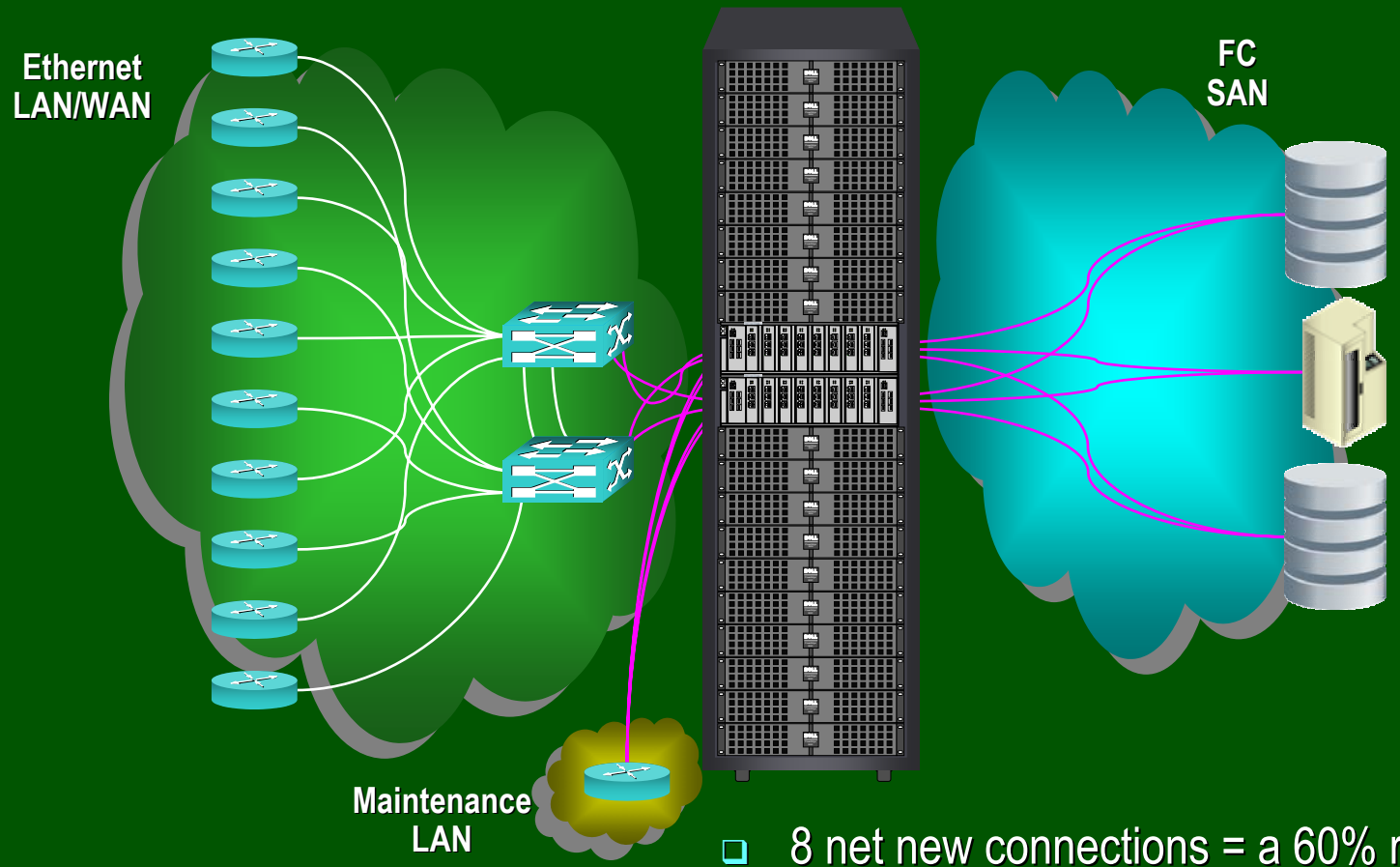
- Adding servers takes a lot of hard work & time



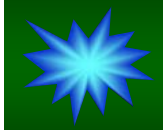
- 20 Net new connections
- Disruptive FC fabric reconfigurations

Managing Scalability w/IB based I/O

- Adding additional servers is significantly simpler & easier

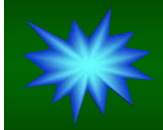


- 8 net new connections = a 60% reduction
  - *(2) IB paths/new server - IB fabric*
  - *No new switches or reconfigurations*
  - *Faster & non-disruptive implementation*



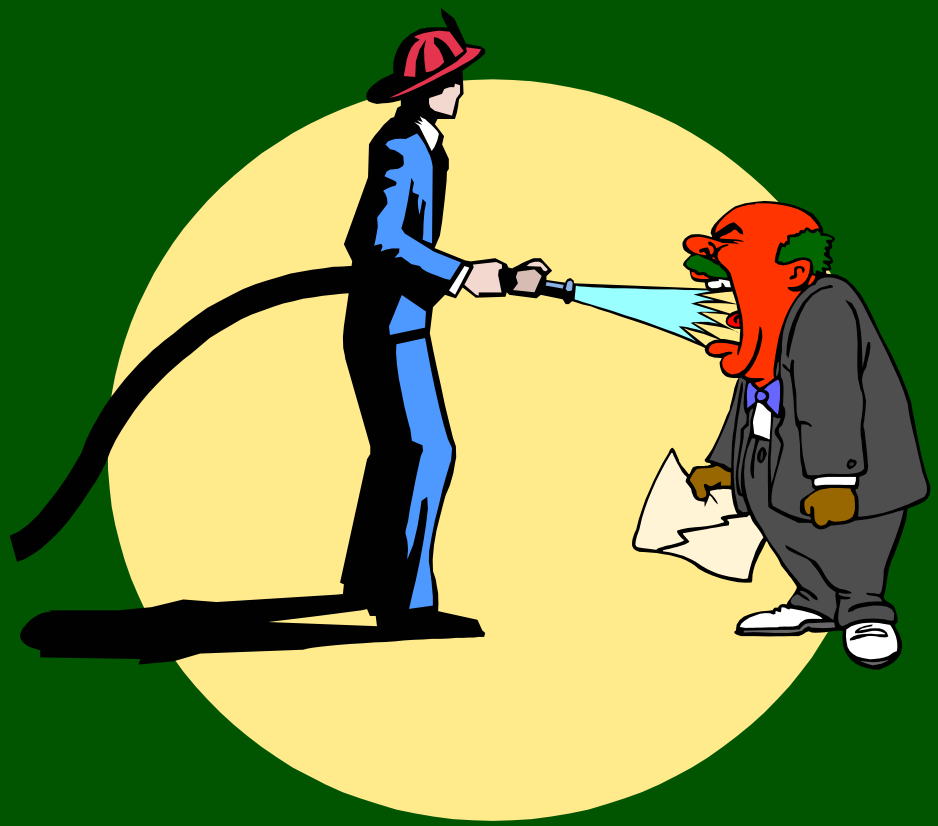
Scalability Net Results w/IB shared I/O

- Making adds, moves, or changes means
  - *Less time*
  - *Less cost*
  - *Less effort*
  - *Less complexity*
  - *Less personnel*
  - *Less disruptions*
  - *More control*
  - *More simplicity*
  - *More stability*
  - *Better RAS*



PCI Bus Constraints

- PCI bus limitations have been strangling CPU I/O
  - *Like trying to drink from a fire hose*



PCI Bus Constraints

PCI	PCI (66Mhz)	PCI-X (133 Mhz)	DDR	QDR	3GIO
Max BW	4 Gbps	8 Gbps	16Gbps	32Gbps	64Gbps
I/O Constraint	(4) GbE w/TOE or (2) 2gig FC	(4) SCSI 320	(1) 4x IB	(1) 10gigE, FC, or 4x IB	(2) 10gigE, FC, or 4x IB
Architecture	Shared Parallel Bus				Switched serial
Issues	Bus contention				Not until 04

PCI Bus vs. IB

Comparison	Advantages	Disadvantages
PCI PCI-X DDR QDR 3GIO	Lower cost Simpler for chip-to-chip Protects software base	Until there is 3GIO, bus contention
InfiniBand	Clustering Scalability: Ports & BW Out-of-box connectivity QoS Security Fault Tolerance Multi-cast Fabric Convergence PCB, Copper, & Fiber	Software

Solution: PCI Bus **AND** IB

- It's not "either:or"
  - *They are complimentary not mutually exclusive*
- The best solutions takes advantage of both
  - *This is why you rarely hear anymore that IB is the PCI replacement*
- There are new HCAs **WITH** PCI-X interfaces
  - *Expect DDR, QDR, & 3GIO as well*
  - *The IB benefits are almost as great*
    - ↪Eliminates bus contention
    - ↪Preserves PCI software base
  - *Provides IB benefits NOW*
    - ↪Don't have to wait for native server IB

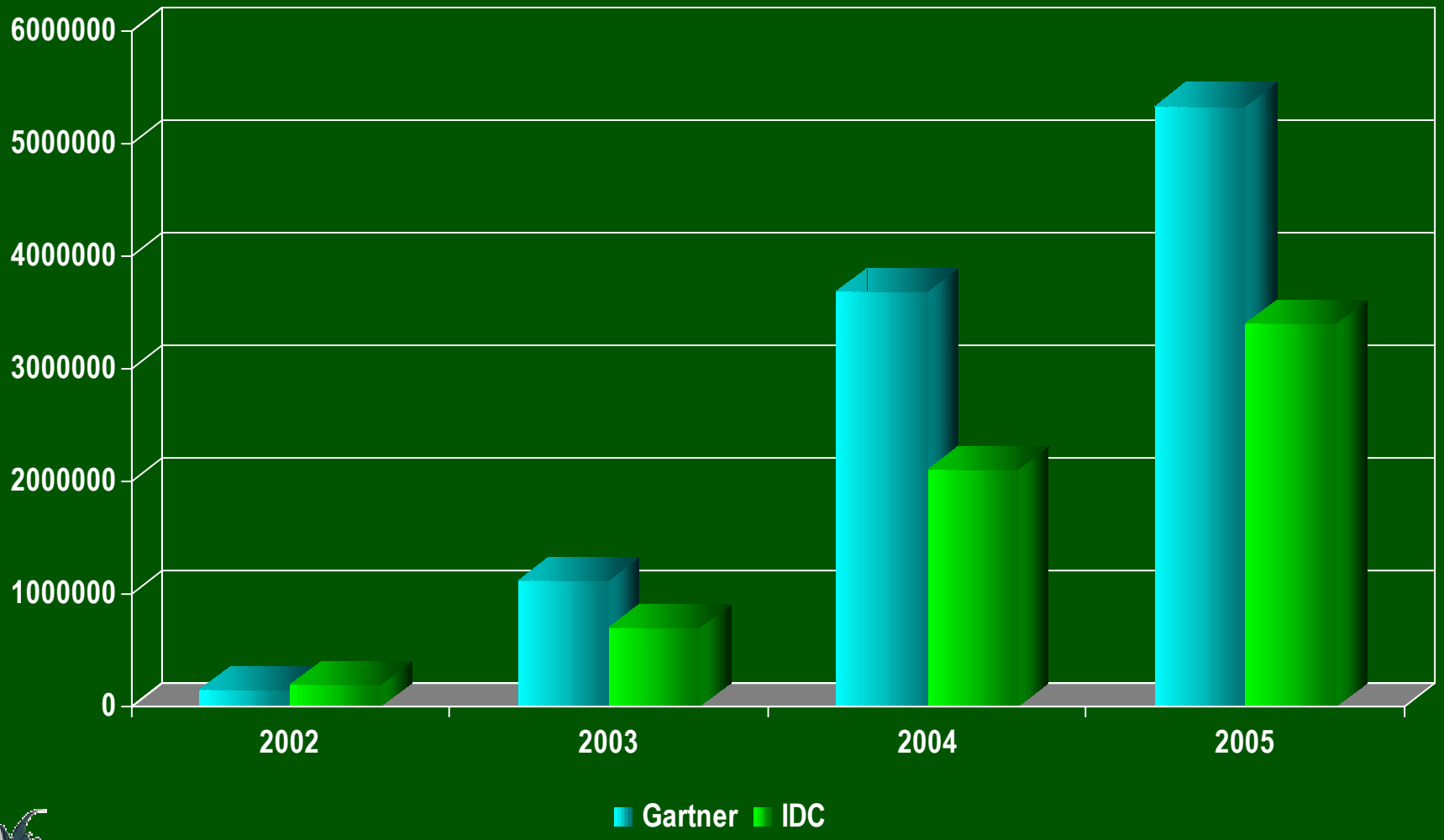
PCI-X HCA Example

Low Cost HP/HA Server Clustering

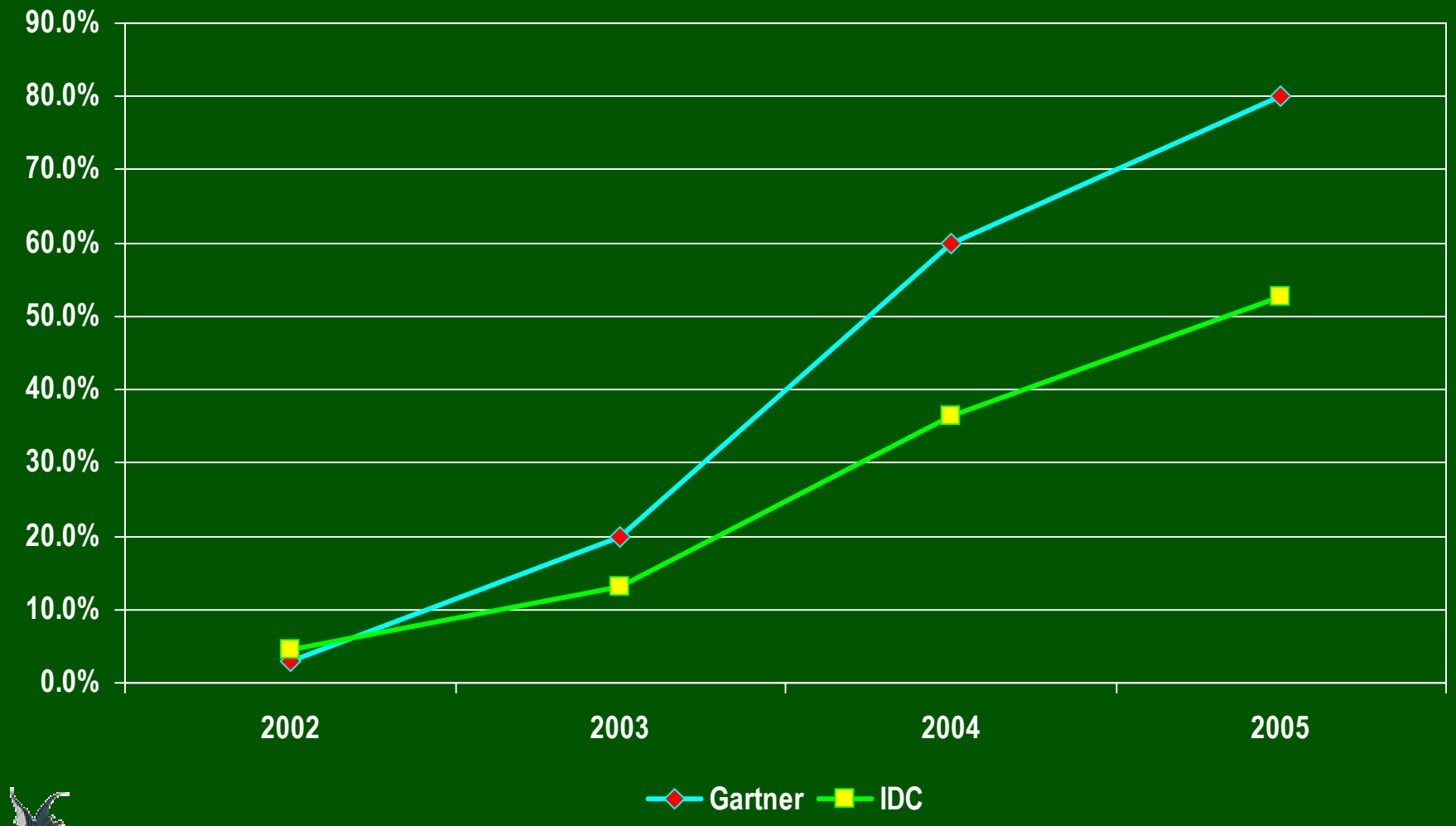
- IB clustering costs less for scaling out than SMP or NUMA scaling up
- IB eliminates fabric messaging performance Issues with clustering
  - *Long queues*
  - *PCI bus contention*
- IB enables low cost server (shared I/O arguments even stronger here)
  - *Diskless blades*
    - Personality on the storage
      - ***Higher Fault Tolerance and Availability***
  - *One connection for clustering and shared I/O*
    - Less I/O interfaces than any other interconnect
      - ***Higher performance***
      - ***Lower TCO***

Industry Analyst's IB Enabled Server Forecast

Analysts are split in their forecast of IB's TAM; but, not on its potential



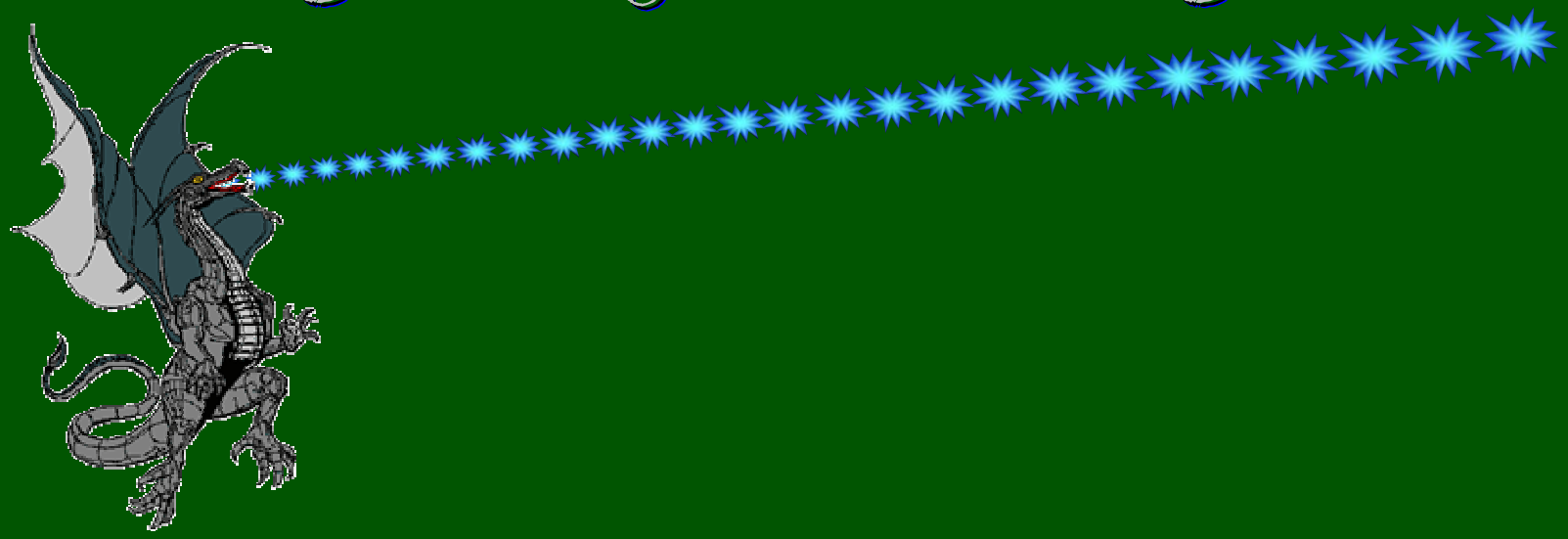
IB Enabled Servers as a % of Total



Conclusions

- Even if the analysts views are optimistic
  - *Huge % of servers will be I/B enabled*
  - *The value proposition is far too strong to ignore*
  - *Initial deployment will utilize PCI-X HCAs*
  - *Native deployments will enable lower cost server blade clusters*
  - *As more and more servers become IB enabled*
    - ➥Clever IT people will realize that they can run IB native for:
      - ➥ ***Clustering, Networking, and Storage***
- When IB becomes native with the server motherboard
  - *The perception becomes that it's free*
    - ➥There is always high market demand for…free.

Dragon Slayer Consulting



?????Questions?????

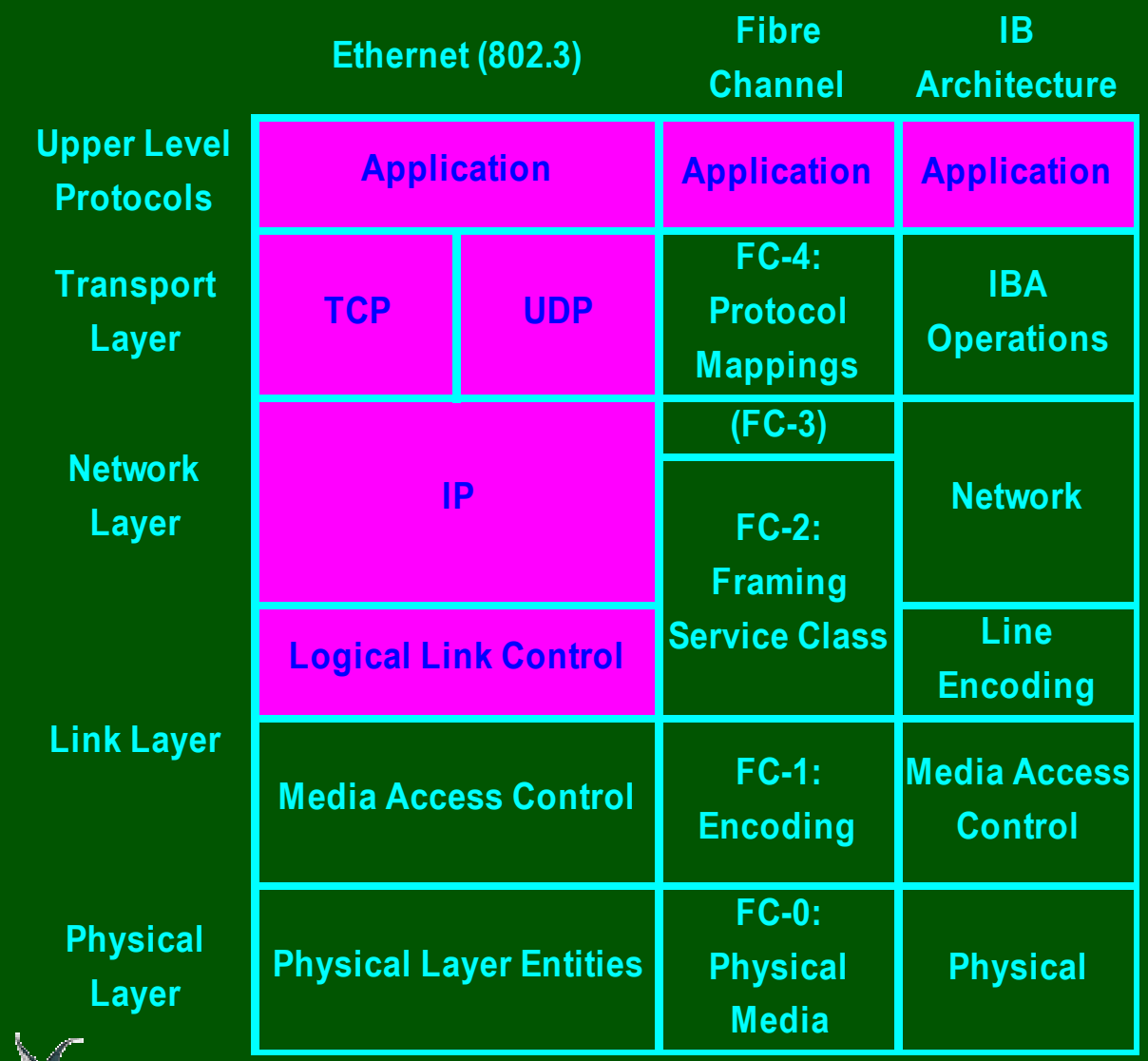
Why Not Just Use GbE or FC?

- GbE and FC are the current fabric infrastructures
- IT personnel already know & understand the technologies
- FC & GbE are already battling it out for SAN infrastructures
  - *FCP vs. iSCSI*

IB vs. FC vs. GbE

Technology	Standards Body	Signaling Speed	First Standard	Maximum Frame Size	Primary Application
Gigabit Ethernet	IEEE & IETF	1.25 Gbps	1999	1.5K	LAN: Local Area Network
Fibre Channel	ANSI	2.125 Gbps	1988	2K	SAN: Storage Area Network
InfiniBand Architecture	InfiniBand Trade Association	2.5Gbps (1x) 10Gbps (4x) 30Gbps (12x)	2001	4K	IAN: I/O Area Network

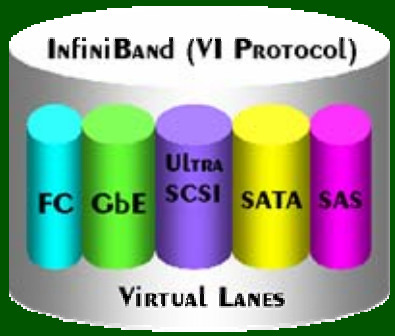
How IB compares w/GbE & FC in OSI



= layers not included in the protocol standards

Data Center Fabric & I/O Consolidation

- IB enables convergence through shared server I/O
  - *One I/O interface for*
    - ➥Clustering
    - ➥Network
    - ➥Storage
      - ➥ ***Eliminates the need for multiple server I/O blades/ports***
  - *IB virtual lanes provides*
    - ➥Multiple independent logical fabrics multiplexed on one physical one
    - ➥QoS to prioritize traffic
    - ➥The benefits of independent fabrics with:
      - ➥ ***The management and maintenance of one fabric***
  - *Switches, directors, and routers provide*
    - ➥Scalability, redundancy, availability, and flexibility



Requirements of a Shared I/O System

- Cooperative Software Architecture
  - *Ability to productively distribute work between host & external shared I/O system*
- Virtualization of I/O
  - *Host manipulates logical resources*
  - *Host has no awareness of underlying physical resources*
- All I/O managed external to host
  - *Host originates requests and receives result*
- Heterogeneous Operating Systems
- 3 Classes of I/O
  - *Efficiently handle small to very large messages*
  - *Microsecond sensitive latency without sacrificing bandwidth*
- Channel Architecture
  - *Highly differentiated priority and service levels*
  - *Connection oriented guaranteed delivery mechanism*
  - *Inherent memory semantics and protection*
  - *High speed / low latency*

Dragon Slayer Consulting

Market Projections

IDC & Gartner-Dataquest