Data Center Ethernet
Facilitation of Enterprise Clustering

David Flynn, Linux Networx

Orlando, Florida

March 16, 2004

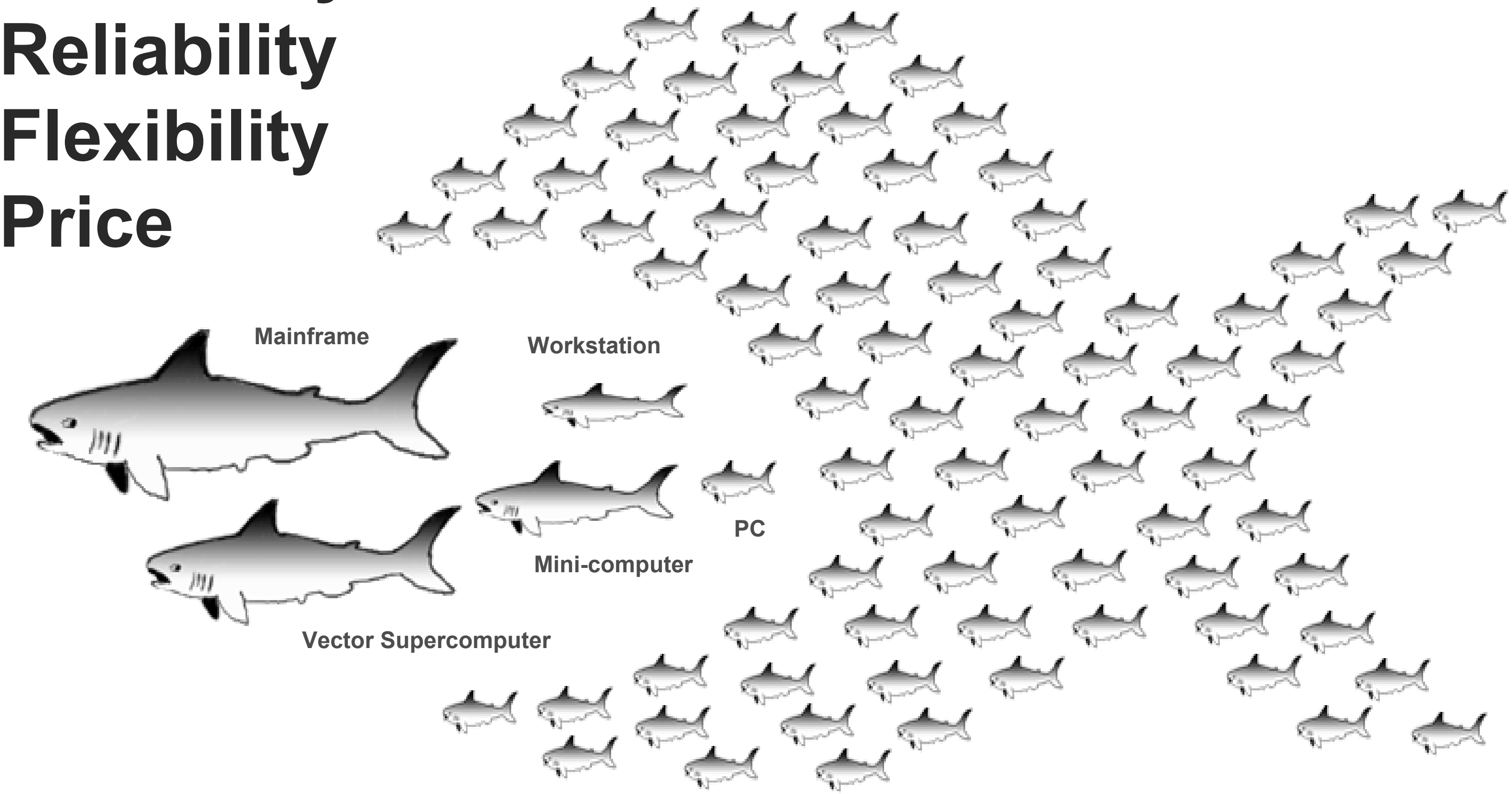
Linux Networx

builds COTS based clusters

Linux Networx™ proven

Clusters Offer Improved

Performance
Scalability
Reliability
Flexibility
Price



Phase-1: Move HPC into Enterprise

6

Los Alamos - Lightning
11.26 TFLOPS
Linux Networx Evolocity
2816 Opteron Processors

7

Lawrence Livermore - MCR
11.2 TFLOPS
Linux Networx E2
2,304 Intel Processors

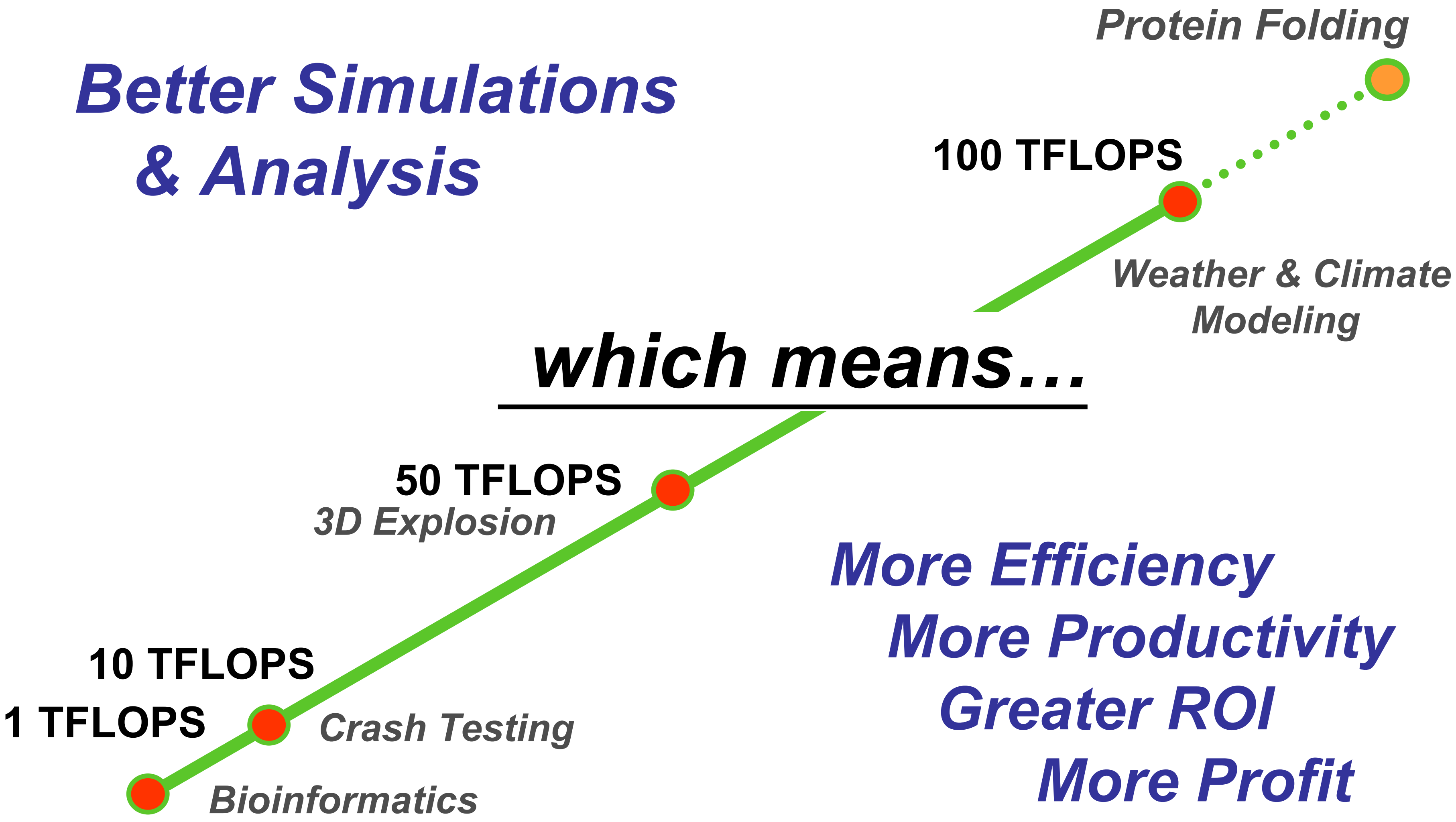
N/A

Los Alamos - Pink
10 TFLOPS
Linux Networx E2
2,048 Intel Processors

LANL BlueSteel – 2.5 TFLOPS, PCR – 1.7 TLOPS,
Argonne – 1.68 TFLOPS, Seitel – 1.4 TFLOPS, Boeing – 1.2 TFLOPS

Cluster Solutions Mean …

Better Simulations
& Analysis



Protein Folding

100 TFLOPS

Weather & Climate Modeling

50 TFLOPS

3D Explosion

10 TFLOPS

1 TFLOPS

Crash Testing

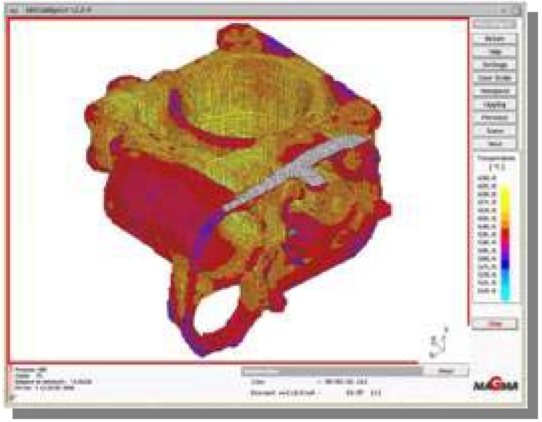
Bioinformatics

which means…

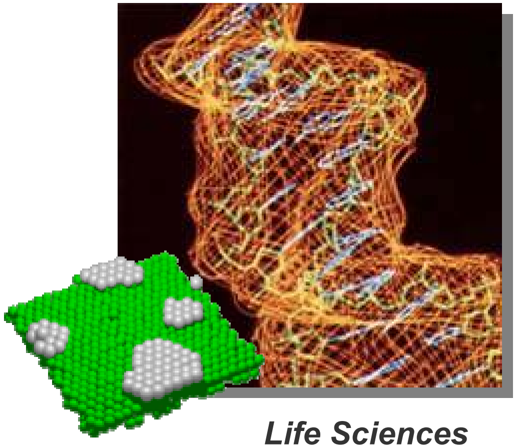
More Efficiency
More Productivity
Greater ROI
More Profit

The HPC Cluster Market

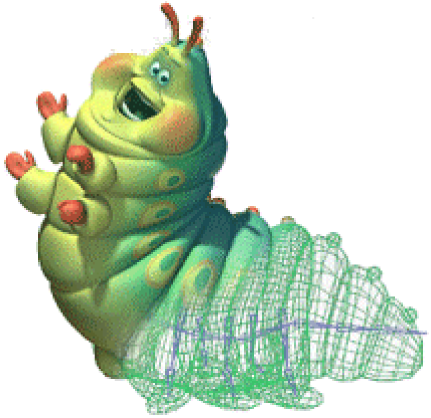
High performance cluster solutions for the scientific, financial & enterprise markets



Engineering Design



Life Sciences



Special FX/Entertainment

Overview	Symbol	Description	Open	Low	High	Last	Bid	Volume	Ask	Source	Bid Size	Ask Size	Dele
	AOL	AOL TIME WARNER INC	[illegible]	[illegible]	[illegible]	[illegible]	0	[illegible]	null	[illegible]	0	null	[illegible]
	CSCO	CSCO SYS INC	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	DELL	DELL COMPUTER CORP	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	T	AT&T CORP	[illegible]	[illegible]	[illegible]	[illegible]	0	[illegible]	null	[illegible]	0	null	[illegible]
	TYC	TYCO INTL LTD NEW	[illegible]	[illegible]	[illegible]	[illegible]	0	[illegible]	null	[illegible]	0	null	[illegible]

Financial Services / Capital Markets

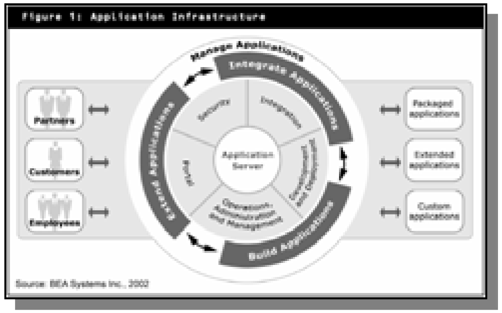
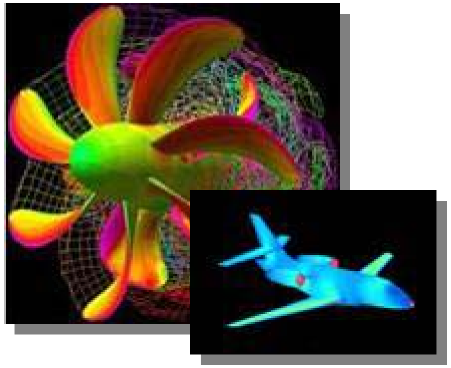


Figure 1: Application Infrastructure

Source: BEA Systems Inc., 2002

Enterprise Infrastructures

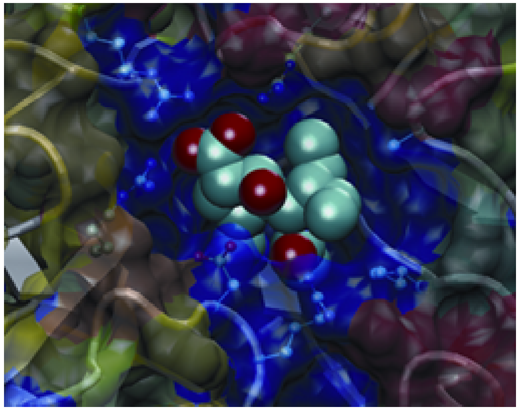


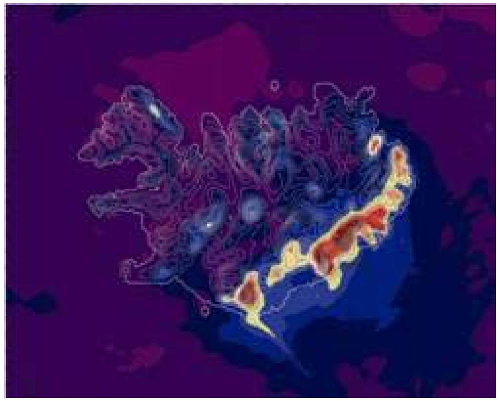
Research & Defense

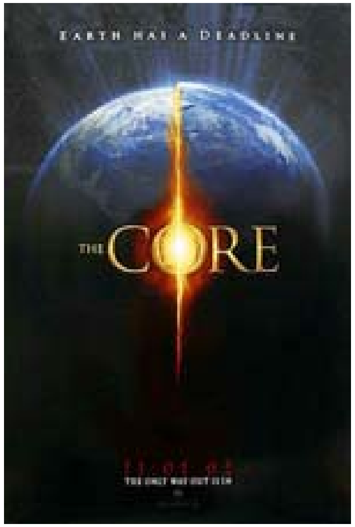
Linux Networx™ proven

HPC Cluster applications

Customers rely on clusters to solve their most complex and technical problems







HPC Clustering in the Enterprise

It's not just a competitive edge

- Faster design and analysis leads to shorter design cycles
- Helps products arrive to market faster
- Ensures safety and effectiveness of products
- Saves time and money

It's required to even compete.

The Expanding Universe of HPC Cluster Independent Software Vendors (ISVs)

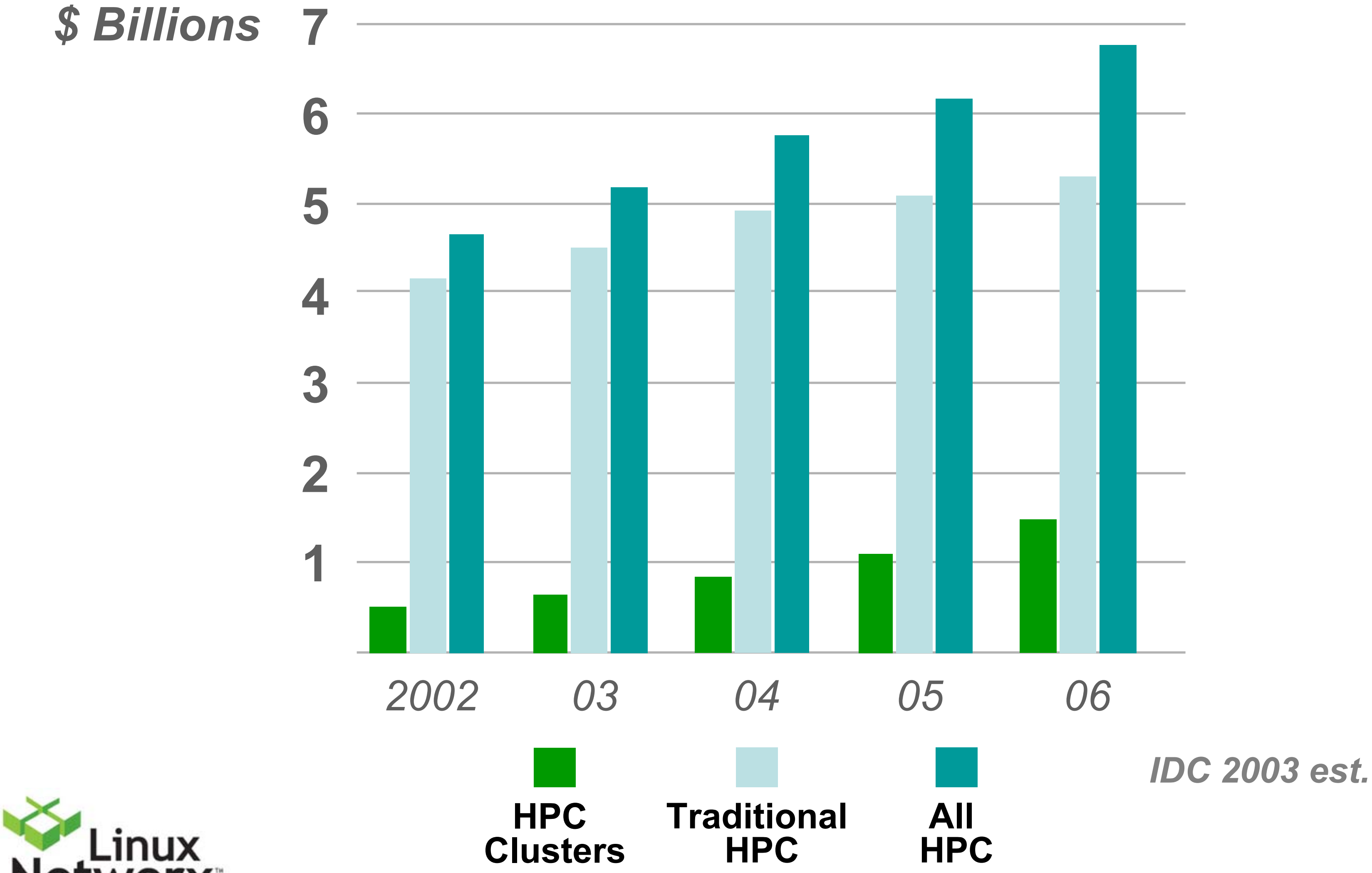
FLUENT
INCORPORATED

ABAQUS
Integrated Technology

MAGMA

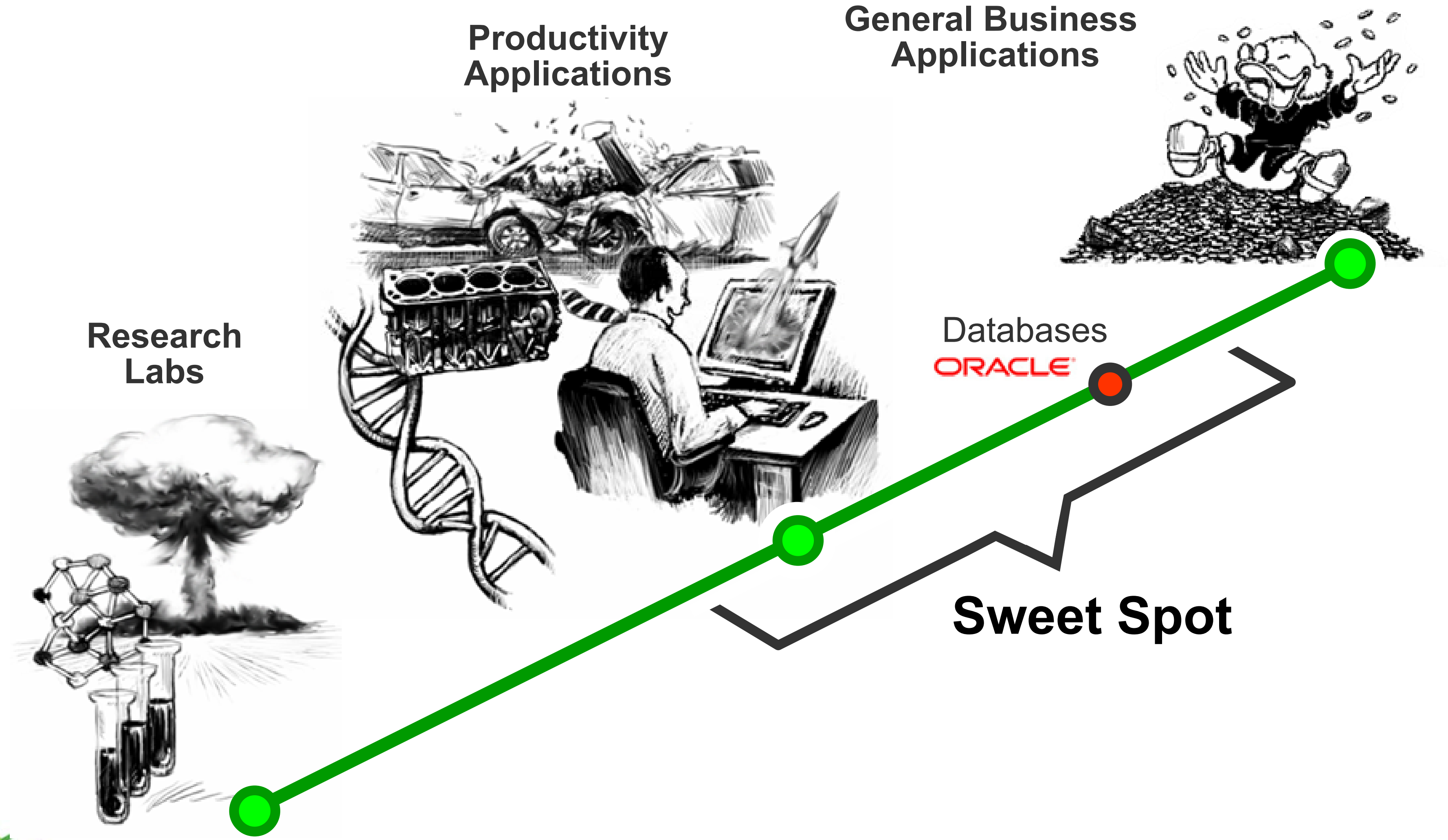
MSC SOFTWARE
SIMULATING REALITY

HPC Cluster Market Growth



IDC 2003 est.

Phase-2: Enterprise Sweet Spot



Research Labs

Productivity Applications

General Business Applications

Databases

ORACLE

Sweet Spot

Linux Networx

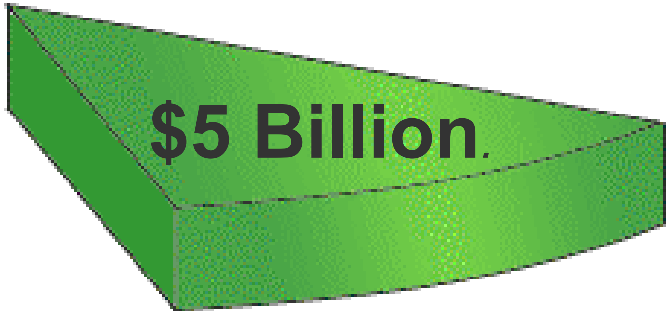
proven

HPC Cluster Technologies Applied to Traditional Enterprise Problems

- Distributed databases
- Distributed transaction processing
- Distributed parallel data mining
- Parallel high performance file systems
- High availability
- N modular redundancy
- Server virtualization
- Utility computing

Pure HPC Market Opportunity

- Scientific Research & Defense
- Engineering
- Oil and Gas
- Life Sciences
- Entertainment

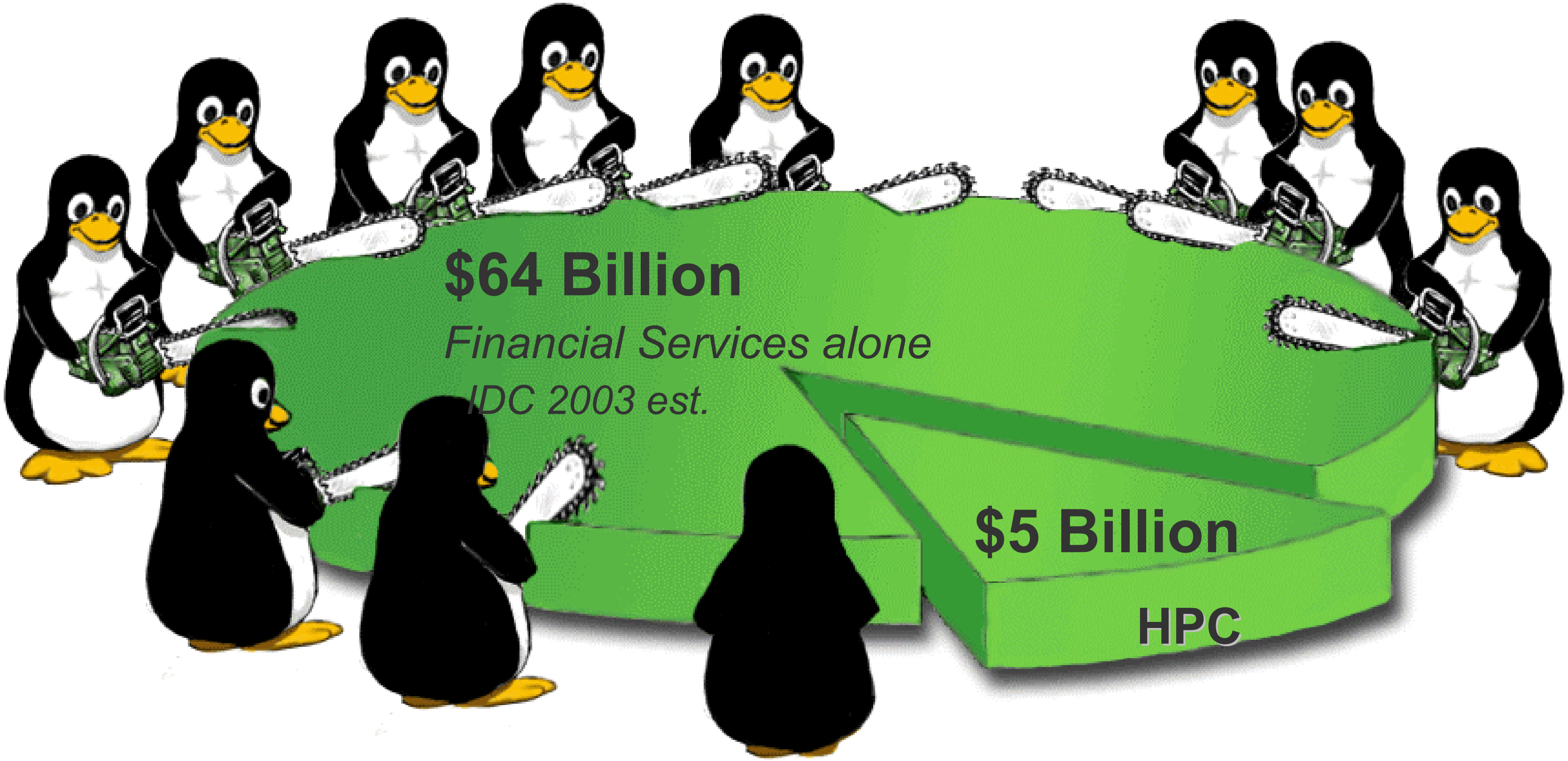


IDC 2003 est.

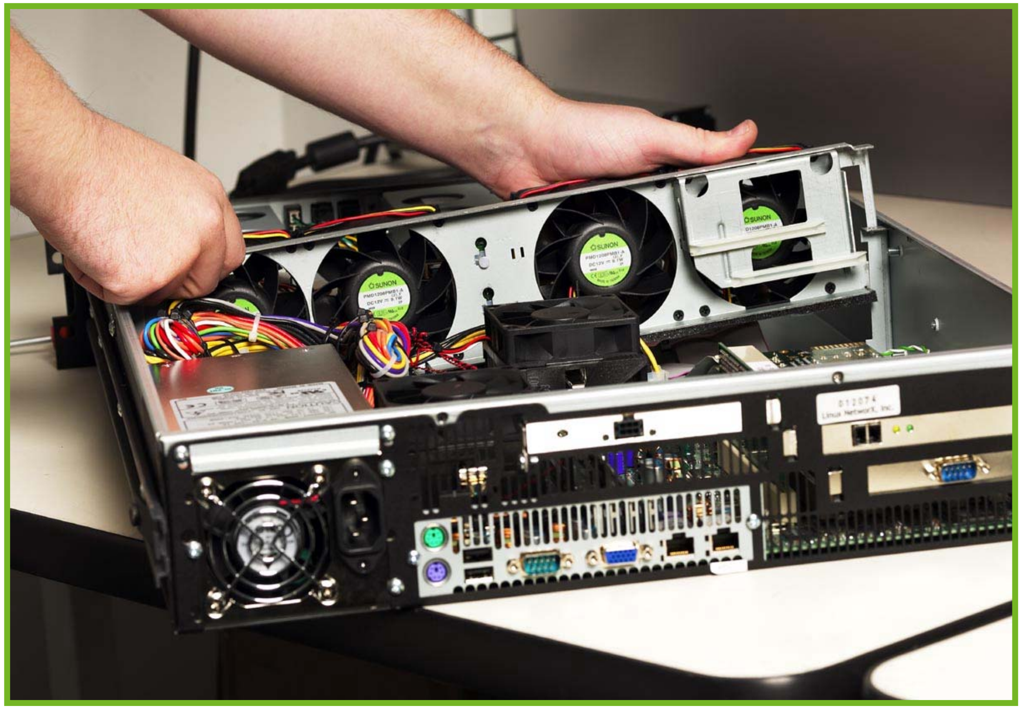
'Future growth is expected to come from all segments of the HPC market'

Combined Market Opportunity

- Financial Services
- Enterprise



What Goes Into a Cluster



- Mother boards
- CPUs
- Memory
- Storage
- Interconnect / network

These must be…

- Based on standards
- Reliable and versatile
- Sold in volume
- Available from multiple vendors
- Richly supported by software and experience
- In a word “commodity”

All can be found COTS - except for the interconnect -

Company 1

Pros

- Most advanced technology
- Designed specifically for HPC
- On-switch barrier & atomic ops
- On-NIC reductions
- 1GB/s bandwidth
- ~3us latency

Cons

- Most expensive
- Single supplier
- Proprietary
- Rigid topology
- Demanding software setup
- Fragile
- Copper only
- No Ethernet interoperability
- Only useful as interconnect
- Needs secondary network
- Limited multicast

Company 2

Pros

- Lowest cost
- Fiber
- On-NIC reductions
- Flexible topologies

Cons

- <260MB/s bandwidth
- ~5us latency
- Single supplier
- Proprietary
- Lack of ECC throughout
- Lack of copper
- No broadcast / multicast
- Limited Ethernet interoperability
- Needs secondary network

InfiniBand

Pros

- Standards based
- Relatively cheap
- Fiber & Copper
- Flexible topologies
- 1GB/s@4X   3GB/s@12X
- DDR 6GB/s @12X
- Link aggregation
- Some Ethernet interoperability
- Multicast

Cons

- ~5us latency
- No on-NIC reductions
- Single supplier for key tech
  - Host client adapters
  - Switches
  - IB to Ethernet bridging
- Too many suppliers of
  - Driver software
  - Management software
- Lacks wide acceptance
- Revolution not evolution
- MAC address incompatibility

What It Takes to Be a Cluster "Interconnect"

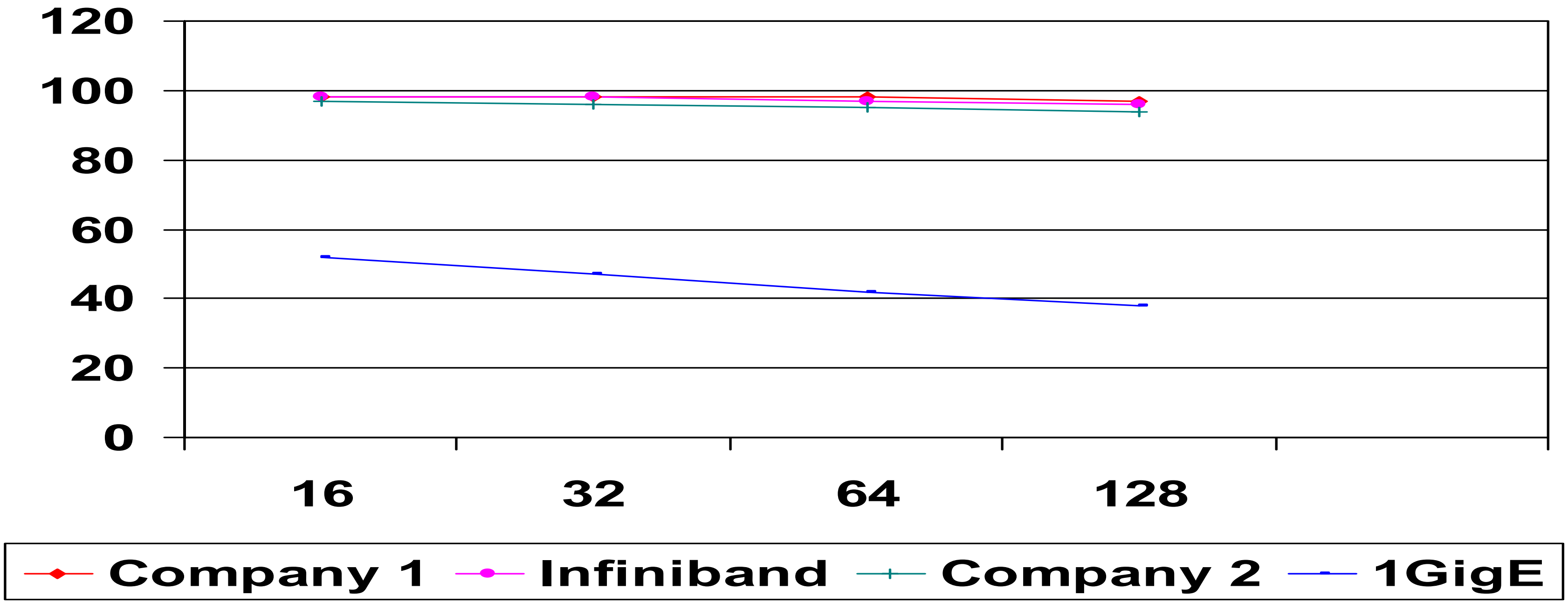
● **Reliable data delivery**	**Got to get the data!**
● **Low latency (<= 3 micro seconds)**	**Got to get it now!**
● **High bandwidth (>= 1GB/s)**	**Got to get lots of it!**
● **Perform well at saturation**	**All need it at once!**

Simulations that run on HPC clusters communicate in bursts. The machines operate in lock step. All compute then all exchange data. The simulation cannot proceed to the next phase until the boundaries are exchanged and communication is complete. Time spent waiting on communication effectively idles the horsepower of the entire cluster.

HPC is then the worst case scenario for network communication in that;
All communicate at once and none proceed until all are done.

Scaling Efficiencies

As a function of interconnect & cluster size



A cluster’s “scaling efficiency” is determined by what percentage of total time the cluster spends calculating.

Time spent communicating is time lost calculating.

Avoiding Priority Inversion

Problem:

- The interconnect must carry traffic of differing priority.
  - High priority synchronization and latency sensitive traffic.
  - Low priority file system access, monitoring and management
- Low priority traffic can be large and slow to be serviced.
- High priority packets may therefore block on large slow packets.
- The cluster nodes stall waiting for important data

Solution:

- The interconnect must implement link level quality of service or QOS.
- Protocols, services, and applications must be coded to operate in a QOS aware fashion.
- Interconnects and HPC software today correctly manage QOS.
- Interconnect today use virtual channels / lanes / networks.

Avoiding Injection Rate Ramping

Problem:

- Ramping up injection rates wastes precious latency/bandwidth.
- Bursting applications cause each other to back off.
- Dropped packets are interpreted as congestion - causing back off.
- Higher bandwidth networks take longer to ramp up.

Solution:

- Use link-level flow control.
- Interconnects today use link credit / debit systems for flow control.
- Multiple buffer spaces or virtual lanes are used to avoid dependency cycles.

Avoiding Packet Loss

Problem:

- Packets are injected on the network at a high rate from many sources.
- These packet storms can overflow switch buffering.
- Packets then get dropped on the floor.
- Data corruption can also cause packet loss.
- Dropped packets must be retransmitted by the sending endpoint.
- Interrupting the sending OS adds significant latency.
- A TOE can help but there is still 2 more network trips of latency.

Solution:

- The interconnect must guarantee delivery of packets.
- Each device ensures uncorrupted deliver to the next.
- Interconnects today guarantee data delivery or are “lossless”.
- Packet drop rates are less than 1 in 1e10th packets.
- Endpoint protocols are “optimistic” – optimized for successful delivery

TCP is Not Suited for Clusters

- TCP is avoided by applications from MPI to Oracle.
- TCP cannot run the network at near saturation.
- TCP cannot even know where saturation is.
- TOEs do not help.
- Pessimistic protocols will not work.
- HPC Interconnects show what will work.
- Optimistic protocols require Data Center Ethernet.
- The MAC must provide key features…

MAC / Ethernet Needs

- Link level QOS
- Link level flow control
- Link level guaranteed delivery
- Protocols and services to match

With these features

- Clustering will migrate smoothly into enterprise.
- Ethernet will become ubiquitous in clustering.
- Clustering will become ubiquitous.

Linux Networx Needs Data Center Ethernet

Enterprise customers want an Ethernet solution

- No one likes having two networks.
  - Doubles the complexity
  - Doubles the maintenance
  - Increased costs
  - Interoperability concerns
- Commodity always wins.
- Everyone knows how to administer Ethernet.
- All software supports Ethernet.
- Proprietary interconnects are the most problematic component.

Linux Networx must provide an Ethernet solution to satisfy enterprise customer requirements.
BONUS: It will save us time and money.