Performance Area Networks

Optimizing High Performance Clusters

Essential Networks

- Corporate history
  - 1983, 1992 Founding
  - $100+ Million Cash, No Debt
  - 2 Founders are technical and long term
  - History of long support for products
  - Only HIPPI switch provider, 75% market share
  - First GSN provider, largest switch, least expensive switch

ESSENTIAL

Corporate Future

- Founders own almost 50% of the shares.
- We're not for sale, we're buying (ideas?)
  - History of successful purchases
  - Successful in changing markets
- Management and investment focus on Essential
- We are committed to extending our lead
  - Investment in products and staff
  - HIPPI innovation
  - Future GSN products
  - Progression to necktop bandwidths

ESSENTIAL

Our Focus on Performance

- Technical leadership
- Long term support
- Proven OEM partner
  - Compaq, HP, IBM, SGI, StorageTek, Sun, Tera (Cray, Inc.)
  - Licensees: Alteon, SGI, Phillips DVS

ESSENTIAL

The Overall PAN Concept

- Definition: A Performance Area Network differs from typical Ethernet LANs in the ability to realistically deliver gigabit and gigabyte throughputs while leaving enough CPU to do more than just data transfer. The goal is to be able to simultaneously process data while importing raw data and exporting results.

ESSENTIAL

OS's We Support

- Beowulf (Linux)
- Compaq Tru64 (Dec Unix)
- HP HP/UX
- IBM AIX
- SGI Irix
- Sun Solaris
- Free and net BSD
- NT, Win2k

ESSENTIAL

Gigabit PAN Candidates

- The goal is high throughput with small CPU loading.

Technology	CPU Usage (Lower is better)	Transfer Rate (Higher is better)
Fast Ethernet	21%	11.5 MB/sec
Gigabit Ethernet	58%	27.3 MB/sec
HIPPI-800	16%	88.1 MB/sec
GSN	19%	560 MB/sec

- With GigE, even 58% CPU loading does not fill even 1/4th of the pipe.
- Note the huge performance advantages of HIPPI and GSN over Gig Ethernet.

ESSENTIAL

PAN Improvements Vary by Environment

Flavor of application	Examples	Issues	PAN Benefits
Batch	Supercomputer jobs, load large image, run, dump	Expensive down time between jobs while loading new input data.	A 5 TB load: 51 hours GigE 16 hours HIPPI 1.8 hours GSN
Flow	Video production, format conversion, telemetry, data reduction	Double whammy. Slow IO plus huge CPU waste on IO. Leaves little CPU..	GigE 27 MB/s 58% HIPPI 88 MB/s 16% GSN 580 MB/s 19%
Backup	Financial, airline, etc.	Limited time period to complete, desire smaller CPU load for backup ops.	1 TB in: 10 hours GigE @ 58% 3 hours HIPPI @ 16% 21 ***minutes*** GSN @19%
Sync	Updates to mirrored web servers, parallel and identical	Simultaneous updates of multiple servers while loaded.	Large updates quickly with negligible performance impact

ESSENTIAL

Finalists in the Oxymoron Olympics

- Giant Shrimp
- Gigabit Ethernet

ESSENTIAL

How much better is a PAN?

- HIPPI compared to Gig Ethernet
  - 300% throughput increase while using less than 1/4 the CPU
  - In flow applications with both large files and high processing (video, etc.), a 10x performance improvement is realistic.
  - 10x (even 3x) is worth much more than the cost of a Performance Area Network

ESSENTIAL

What does this mean?

- Platform vendors don't properly position PAN solutions.
  - Like putting bicycle tires and a motorcycle transmission on a new Ferrari
  - Put a 10x performance penalty on each new system
  - Its getting worse… IO bandwidths are not keeping pace with Moore's law bandwidth improvements apart from GSN and HIPPI.
  - Sales of $500,000+ computing systems on Ethernet should be prosecuted as a felony.

ESSENTIAL

Conclusions on Performance

- We in the HNF need to carefully explain:
  - costs vs. performance gains:
    - faster processors
    - more memory, etc.
  - your biggest bang per buck is on a PAN.
  - On any cluster that processes large source files --you can get 3x - 10x improvement with minor expenditures.
- The question is how to explain it to management (or customers).

The ROI Model
(how engineers can talk to bean counters)

- All accountants and planners will approve any expenditure that save more money than it costs. (The one year payoff approach.)
- CFO types will also approve any project that exceeds the ROI (return on investment) average of the corporation.
- Everything above 18% ROI is approved.

ESSENTIAL

ROI Model Example

- Start with your machine cost per hour
- Then calculate time not computing due to IO operations.

Machine	Staff	Overhead	Annual	Per Hour				
$1,000,000	$1,000,000	$150,000	$1,483,333	187.29				
5TB Loads	Load Time	Job CPU	Jobs per	Hours per	Days Blown			
	(hours)	(hours)	year:	job	Loading data			
100BT	510	72	14	582	297			
Gig Eth	51	72	64	123	136			
HIPPI	16	72	90	88	60			
GSN	2	72	107	74	9			

ESSENTIAL

ROI Calculations

- Upgrading to a PAN buys extra productive days per year:

	Annual Days Saved changing to	Annual Days Saved changing to	Annual Days Saved changing to	
From	GigE	HIPPI	GSN	
100BT	162	238	289	
Gig E		76	127	
HIPPI			51	

ESSENTIAL

ROI Calculations

The “value” of previously wasted CPU hours gained back by performance enhancements:

	New	New	New			
Old	GigE	HIPPI	GSN			
100BT	$1,151,831	$1,252,593	$1,288,927			
Gig E		$428,519	$595,955			
HIPPI			$235,610			

ESSENTIAL

The Actual ROI for Upgrades

- Your accountant or budget director will appreciate ROI numbers.
- The increase in the number of jobs per year may also save purchasing multiple systems, which has not been included here!
- Also worth calculating is the increase in CPU % available for job processing as a function of the higher efficiency of PANs, which can easily 2x - 4x the jobs processed per year numbers again!
- My example does not consider the 40% CPU boost either.

	New	New	New	
Old	GigE	HIPPI	GSN	
100BT	1280%	746%	96%	
Gig E		255%	44%	
HIPPI			18%	

ESSENTIAL

GSN?

- *Gigabyte System Network*
  - *800 MB per Second*
  - *Error Free*
  - *Flow Controlled*
  - *Highest Bandwidth*
  - *Lowest Latency*
  - *Compatible with HIPPI and GigE*

ESSENTIAL

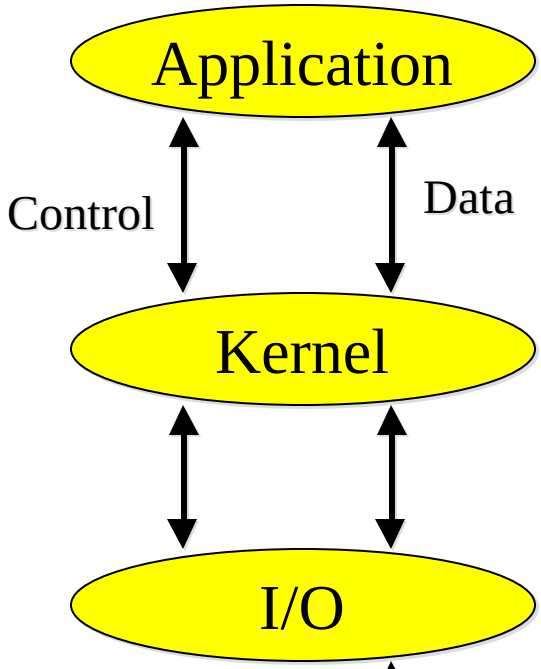
The Main Performance Goals of GSN

- The next standards-based, performance leap
- Extremely low latency
- Compatibility with existing technologies
- Error-Free, Flow controlled links
- Support for OS Bypass
- Interleaving of multiple messages on a single link

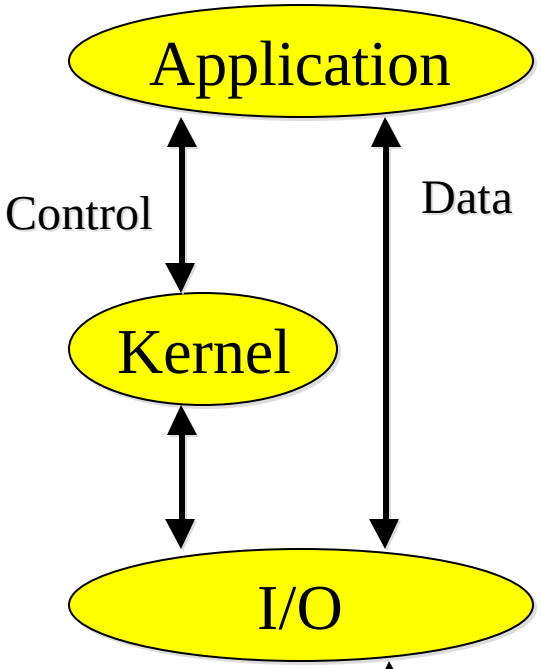
ESSENTIAL

Scheduled Transfer: Bypassing OS Structures

Allows for: Low Latency to End User, Low CPU Usage, and Reduced Network Congestion



Traditional Approach:
Data and Control
Through Kernel



OS Bypass:
Data Isn't Processed by
OS Kernel

ESSENTIAL

How ST works

- Both Endpoints are prepared for data movement before data is transmitted
- VC is established
- Handshake for memory allocation
  - Single Use
  - Multiple Use
- ST Control completes pre-arrangements
- Data moves
- Last packet info is sent on tailbit of final data micropacket

The Essential 10000

ESN-10000 Features:

- Only 32 port GSN Switch
- 32 full duplex 8 Gigabit/s per port = 512 Gbps
- 14 GSN Systems Delivered
- Protocols Supported:
  - HIPPI-6400-PH (Physical Layer)
  - HIPPI-6400-SC (Switch Control)
  - ST (Scheduled Transfer)
- Hot swappable Ports, Redundant Power Supplies
- Only full featured little brother..

ESSENTIAL

GSN at LANL

ESSENTIAL

The Essential 8000

- Announced and shown at SuperComputing Portland 11/99
- Lowest cost 8 Port GSN Switch.
- Only GSN family with both an 8 port entry model and a guaranteed migration strategy to a production size 32 port GSN switch.
- Real GSN products, field proven, history of quality engineering and support.
- What's next?

ESSENTIAL

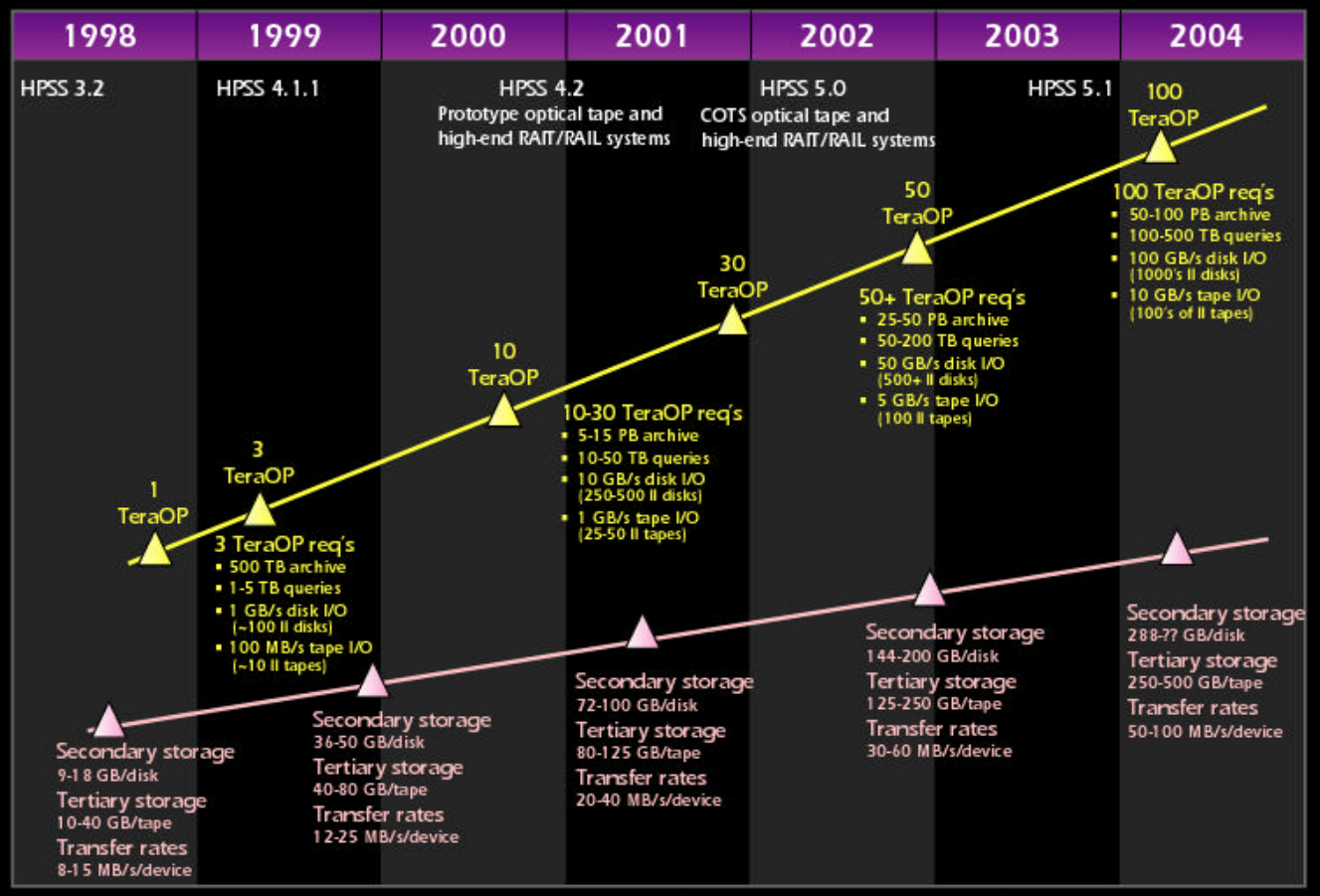
OS Bypass

- Little packets meet fast processors, processors lose
- Inefficiency of host TCP stacks
- Huge and increasing CPU burden for IO
- We’re starving our processors and its getting worse, Ethernet and 1500 byte packet sizes processed by the host are killing us.
- The old model for HPC clusters is dead
- ST support on ever-faster Layer 2 products are required to keep up with Dr. Moore after 2004.

ESSENTIAL

High-performance computing

ASCI storage requirements driving HPSS and the need for high-capacity, scalable devices



http://www.llnl.gov/asci/

ESSENTIAL

Current GSN Performance

- All tests are on IRIX 6.5.6f, 250 MHz R10000 Origin 2000 with 8 CPUs between two idle systems.
- GSN support for Onyx2 now available.
- Sockets API

Protocol	Bandwidth MB/sec	CPU Utilization percent of a single CPU.
Sockets - STP	560	16% Tx, 21% Rx
Sockets - UDP	473	145% Tx, 105% Rx
Sockets - TCP	250	100% Tx, 100% Rx

- OS Bypass API (libst)
  Latency 6 uSec one-way, 64 byte user payload, user process to user process
  Bandwidth 538 MB/sec using 16 KB messages, single producer/consumer
  Packets/sec 1.5 million 64 byte user payload packets per second
- Everybody uses ST on GSN, so you get 1000% higher transfer rates compared to Gig Ethernet and 700% higher performance than HIPPI.

ESSENTIAL

The Steep Trail and the holy grail

- Acceleration of Moore's law since 1980, 14.3 months now (delta -0.216 mo./yr.)
- % Necktop Performance by year
  - Human 1^{16} C/s
  - When is human parity year for: Cluster, SC, PC?
- IO Bandwidth by year, by machine type
- We won't get there without ST and specialty IO processing in the legacy of GSN. Time to pioneer is now. It's a steep trail ahead.

ESSENTIAL

Expected Capabilities and IO Requirements through 2023

					#NHB in a	#NHB in a			
	A $1000 PC		Server IO	WS IO	2000 Node	20000 Node	SuperComputer		
	Calc/second		Gb/s	Gb/s	Cluster	Cluster	C/s		
2000	8.29E+08	Insect 1e8	0.829	0.083	0.0002	0.002	5.029E+12	Mouse 1e11	
2001	1.41E+09		1	0.141	0.0003	0.003	8.56E+12		
2002	2.44E+09		2	0.244	0.0005	0.005	1.48E+13		
2003	4.28E+09		4	0.428	0.0009	0.01	2.599E+13		
2004	7.64E+09		8	0.764	0.002	0.02	1E+14		
2005	1.39E+10		14	1	0.003	0.03	1.814E+14		
2006	2.56E+10		26	3	0.005	0.05	3.344E+14		
2007	4.79E+10		48	5	0.01	0.1	6.27E+14		
2008	9.14E+10		91	9	0.018	0.18	1.196E+15		
2009	1.77E+11	Mouse 1e11	177	18	0.035	0.35	2.321E+15		
2010	3.5E+11		350	35	0.07	0.7	4.584E+15		
2011	7.05E+11		705	70	0.14	1.4	9.22E+15		
2012	1.44E+12		1443	144	0.29	2.9	1.889E+16	Human 1e16	
2013	3.01E+12		3012	301	0.6	6	3.941E+16		
2014	6.41E+12		6408	641	1	10	8.383E+16		
2015	1.39E+13		13896	1390	3	30	1.818E+17		
2016	3.07E+13		30738	3074	6	60	4.021E+17		
2017	6.94E+13		69376	6938	14	140	9.077E+17		
2018	1.6E+14		159843	15984	32	320	2.091E+18		
2019	3.76E+14		376101	37610	75	750	4.921E+18		
2020	9.04E+14		904150	90415	181	1810	1.183E+19		
2021	2.22E+15		2221802	222180	444	4440	2.907E+19		
2022	5.58E+15		5583586	558359	1117	11170	7.305E+19		
2023	1.44E+16	Human 1e16	14357794	1435779	2872	28720	1.878E+20		

(GSN Satisfies Boxes, the proposed 12000 extends this.)

ESSENTIAL

Contact Us

- Copy of presentation
  Joe Head - head@ods.com
- John Freisinger - johnf@ods.com
- Local support: Paris, Munich, London
  Emmanuel Trublereau emmanuel@ods.com
- Nova Systems: France, Bernard Morin
  Geneva, Arnaud St. Giron
- Invitation to visit us: Albuquerque and Dallas
- Our other business, network security:
  www.intrusion.com

ESSENTIAL