CAC 2007 Panel Position:

Accelerators In Cluster Communication

Greg Pfister
Distinguished Engineer
IBM Systems and Technology Group
Austin, TX

ON DEMAND BUSINESS™

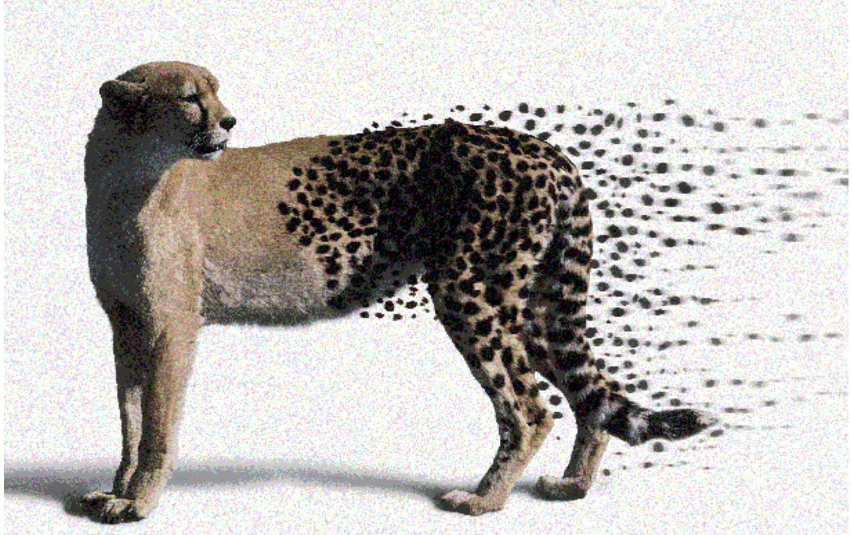
The Legal Page

Any relationship between the opinions expressed here and any official position of the IBM Corporation…

- or of anyone else in the IBM Corporation

...is purely coincidental.

And all copyrighted words are the property of their respective owners.



Random gratuitous clipart

The Questions, and Some of My Answers

Will full or partial offload be the norm for Ethernet?	Partial
Will any collective functions ***typically*** be offloaded to the NIC in the near future?	No
What about collective support in switches **{*typically* }**?	Heck no
Should NICs assist in "read-modify-write" capability to support remote updates?	No, but.
Required	
What communication functions should **always** be offloaded to accelerators?	later
Which should **never** be offloaded to accelerators?	later
Are there cases for which it makes sense to offload to another core rather than to an accelerator? (For SMT cores, I suppose one might also consider offloading to other threads of the same core?)	later, but of course.

Sample of Industry Acceleration Activity

XML, Java	XML Processing	Actional, Cisco, Conformative, DataPower, Reactivity, Sarvega, Tarari
	Java Virtual Machine	Azul
Networks: Storage, Cluster, LAN	TCP/IP + Ethernet	Adaptec, Allied Telesyn, Amasso, Brocade, Chelsio, Cisco(Topspin), Citrix, Crescendo Networks, Enigma Semiconductor, Infrant, NetEffect, NextIO, Nortel, Precision I/O, Silverback, Sensory Networks, Tehuti, Toplayer , Voltaire
	InfiniBand	Cisco (Topspin), Mellanox, SilverStorm, Voltaire
Application Acceleration		Cisco(AON), Citrix, F5 Networks, Juniper Networks(Redline), Radware
Security, Privacy, Rights Management	Trusted Environment	STMicroelectronics, Intel, Infineon
	Cryptographic Functions	Actional, Allied Telesyn, Broadcom, Check Point, 3Com, Cipheroptics, Cisco, DataPower, Enterasys, Forum Systems, HP, Intel, Kasten Chase, Lucent, nCipher, Nokia, Nortel, F5 Networks, Radware, Reactivity, Sonicwall, Sun, Tarai, Vormetric
Real-time Analytics	Data Warehouse Based	Cogent Systems, DATAllegro, Netezza
	Data Stream Based	Tarari
Collaboration & Information Mgmt	Audio, video, data, IM Fusion	ClearSpeed, Tarari
	Media Distribution	Tarari, SeaChange
	Enterprise Search	Netapp, Google, Search Cacher, Thunderstone
High Performance Computing	Floating Point & Integer Comp	ClearSpeed, Cray, Mitrionics, SRC, Tarari, TimeLogic
	Messaging	Cisco (Topspin), Mellanox, Myricom, Quadrics, QLogic, Voltaire, …
	Reconfigurable Systems & SW	Cray, SRC, Nallatech, Tarari, Cadence, Mitrion, Koan, Synopsys, Celoxica, Impulse, Starbridge, SGI
Intelligent Storage Network	Storage Virtualization	Acopia, Brocade, EMC, HP, Hitachi, Index Engines, IBM, NeoPath, Sun, Troika,
	Storage Services	
Accelerator Technology	System-on-Chip	Arteris S.A., Bay Microsystems, Broadcom, Cavium, Freescale Semiconductor, Infineon, LSI Logic, Rapport (Kilocore), Raza Microelectronics, STMicroeletronics, Teja, Tensilica
	ASIC	Advanced Architectures, Britestream, Cavium, Critical Blue, Elixent, Forte, Freescale Semiconductor (Seaway Networks), IP Fabrics, LSI Logic, Mellanox, nCipher, Propulsion Networks, STMicroeletronics, Xelerated
	FPGAs	Xilinx, Altera, Lattice, Acte

Sample of Industry Acceleration Activity

XML, Java	XML Processing	Actional, Cisco, Conformative, DataPower, Reactivity, Sarvega, Tarari
	Java Virtual Machine	Azul
Networks: Storage, Cluster, LAN	TCP/IP + Ethernet	Ac[illegible] Al[illegible] T[illegible] [illegible]asso, Brocade, Chelsio, Cisco(Topspin), Citrix, Crescendo Networks, Enigma Semiconductor, Infrant, NetEffect, NextIO, Nortel, Precision I/O, Silverback, Sensory Networks, Tehuti, Toplayer , Voltaire
	InfiniBand	Cisco (Topspin), Mellanox, SilverStorm, Voltaire
Application Acceleration	[illegible]	[illegible]itrix, F5 Networks, Juniper Networks(Redline), Radware
Security, Privacy, Rights Management	Trusted Environment	STMicroelectronics, Intel, Infineon
	Cryptographic Functions	Actional, Allied Telesyn, Broadcom, Check Point, 3Com, Cipheroptics, Cisco, DataPower, Enterasys, Forum Systems, HP, Intel, Kasten Chase, Lucent, nCipher, Nokia, Nortel, F5 Networks, Radware, Reactivity, Sonicwall, Sun, Tarari, Vormetric
Real-time Analytics	Data Warehouse Based	Cogent Systems, DATAllegro, Netezza
	Data Stream Based	Tarari
Collaboration & Information Mgmt	Audio, video, data, IM Fusion	ClearSpeed, Tarari
	[illegible]	[illegible]
	Enterprise Search	Netapp, Google, Search Cacher, Thunderstone
High Performance Computing	[illegible]	[illegible]TimeLogic
	Messaging	Cisco (Topspin), Mellanox, Myricom, Quadrics, QLogic, Voltaire, …
	Reconfigurable Systems & [illegible]	[illegible] Cadence, Mitrion, Koan, Synopsys, Celoxica, Impulse, Starbridge, SGI
Intelligent Storage Network	[illegible]	[illegible]th, Sun, Troika,
	Storage Services	[illegible]
Accelerator Technology	System-on-Chip	A[illegible] Ba[illegible]sys, Broadcom, Cavium, Freescale Semiconductor, Infineon, LSI Logic, Rapport (Kilocore), Raza Microelectronics, STMicroelectronics, Teja, Tensilica
	ASIC	Advanced Architectures, Britestream, Cavium, Critical Blue, Elixent, Forte, Freescale Semiconductor (Seaway Networks), IP Fabrics, LSI Logic, Mellanox, nCipher, Propulsion Networks, STMicroeletronics, Xelerated
	FPGAs	Xilinx, Altera, Lattice, Acte

This is a zoo

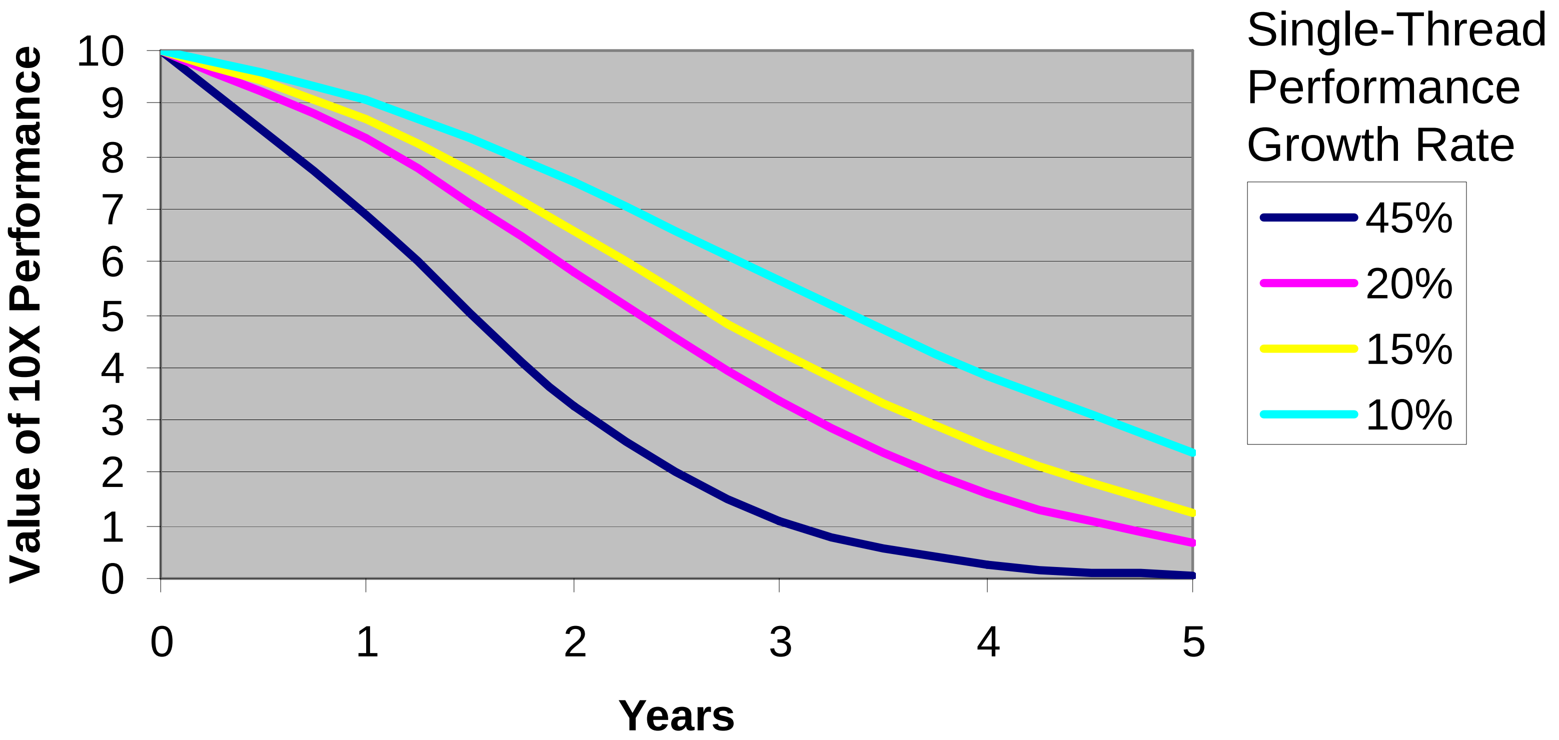
- There is no single accelerator architecture.
  - Some are totally unique hardware.
  - Many are general-purpose systems with a special algorithm, security chip, NIC, etc.
- There is no single accelerator programming model.
  - Subroutine, coroutine, SOA, SQL, passthrough, shared memory, message-based, system access, user-mode access, …
- There is no single accelerator attachment technique.
  - Ethernet, IO bus adapter, part of chipset, etc.

The strategy of acceleration is: *specialize to the application*.

- To optimize performance, power, area, cost

- The tactics – designs – are completely variable.

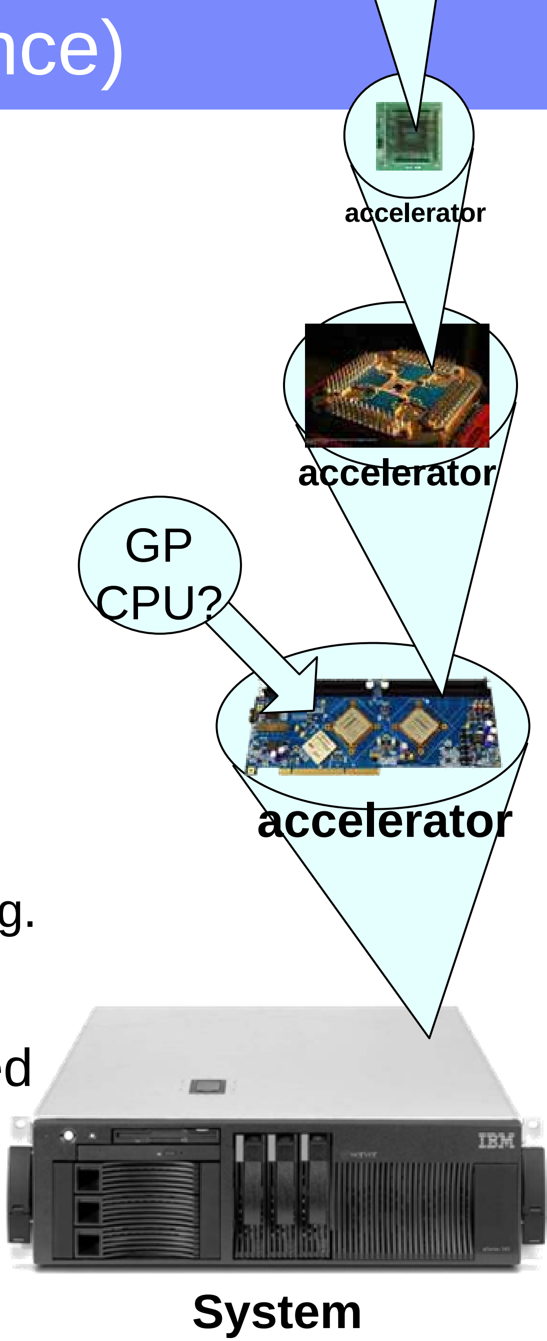
Accelerator Longevity (Controversial!)



- Frequency growth slowdown ⇨ enhanced business case:
  - Past (45%): After 3 years – useless.
  - Now (<20%): as much 5 years useful lifetime
- Controversy? *Aggregate* chip performance still at 45%
  - There will be lots of cores and threads.
  - Must use them all (nontrivial), but accelerators often parallel too.

What is an Accelerator? (vs. an Appliance)

- **NOT a general-purpose system.**
  - ƒ The strategy is specialization
  - ƒ “GP accelerator” is an oxymoron
- **Accelerator:** a device optimized to enhance the performance or function of a computing system.
  - ƒ *Does not function on its own.*
    - n Requires invocation from host programs.
    - n Intention & design optimization, not physics
    - n May contain GP system parts (like CPU) and be substantially software or firmware.
  - ƒ May contain other accelerators
    - n E.g., crypto in protocol offload in XML processing.
- **Appliance**: a device that performs a complete customer-visible function with substantially reduced management & programming requirements.
  - ƒ Functions on its own.

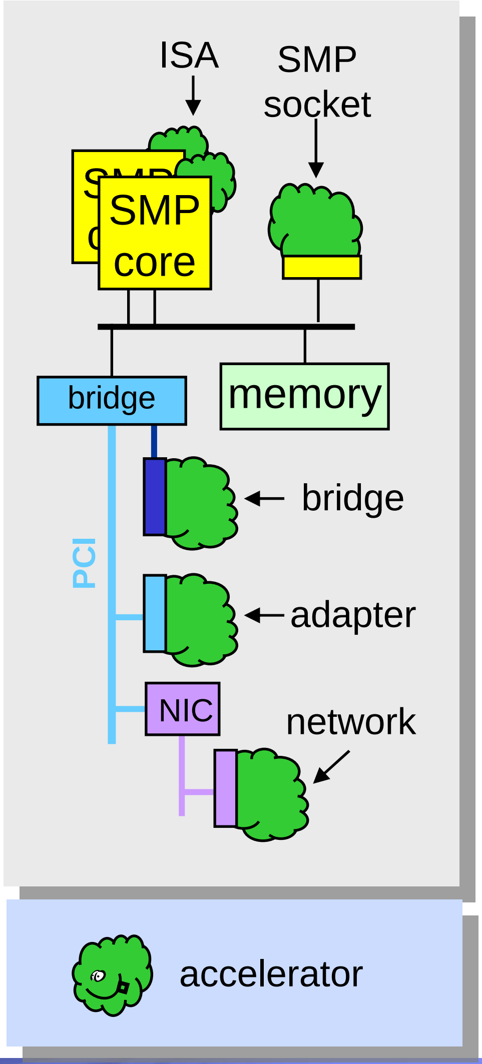


Must **Attach** to a Computing System

- Accelerators have been attached to every conceivable orifice of a computer

Tighter coupling
Lower overhead
Faster synchro.
Integrated SW
Broader use

Narrower use
Slower synchro.
Higher overhead
Loose coupling

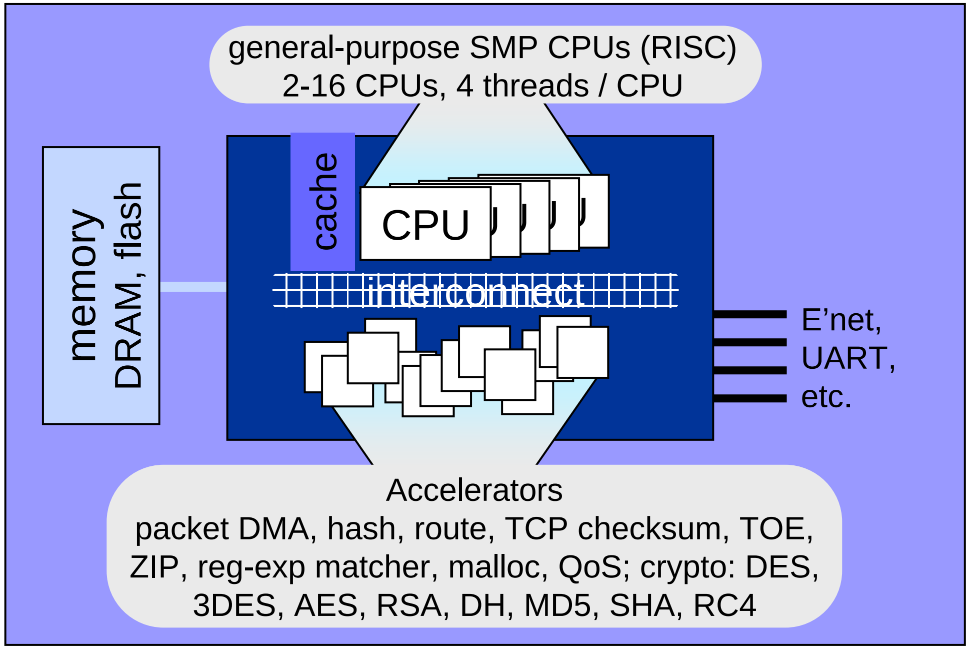


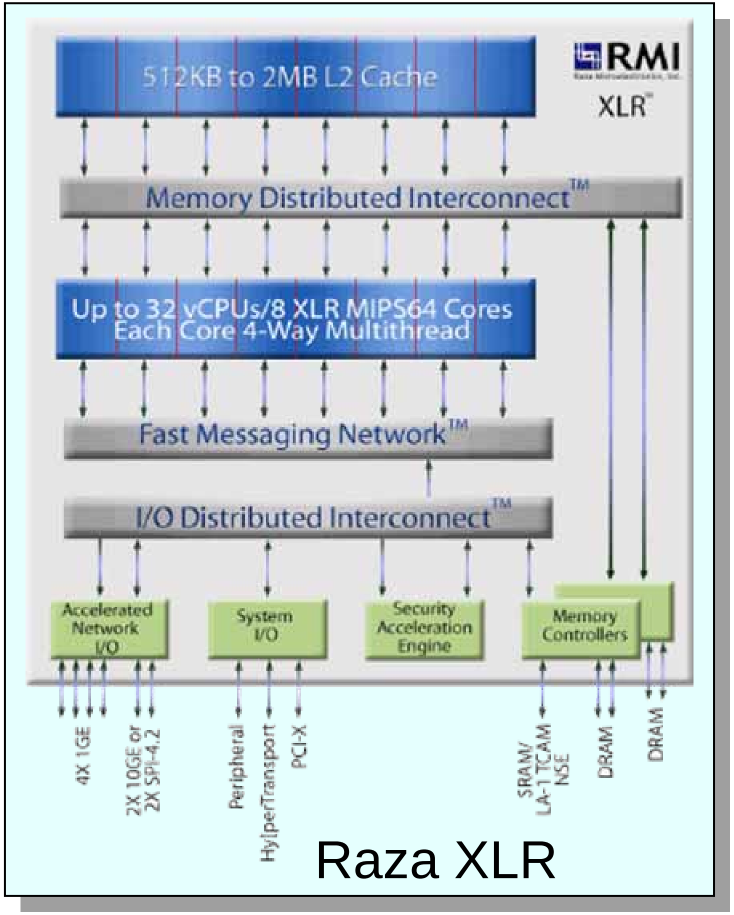
High dev expense
Arch. dependent
Longer lead times
High SW expense
3rd party⇒lower RAS
Few entrants
(AMD: 4; Intel: 2)

Many entrants
Lower SW exp
Fast time-to-mkt
Arch. independent
Low dev expense

(memory attach, too – proposals)

Network Sys/chip: Arms Dealers to Appliances/Accelerators


general-purpose SMP CPUs (RISC)
2-16 CPUs, 4 threads / CPU
memory
DRAM, flash
cache
CPU
interconnect
E'net,
UART,
etc.
Accelerators
packet DMA, hash, route, TCP checksum, TOE,
ZIP, reg-exp matcher, malloc, QoS; crypto: DES,
3DES, AES, RSA, DH, MD5, SHA, RC4



- Values: low power, size, cost, parts count
  - Allow appliance/accelerator focus on new value add, not standard stuff
  - Uses: SSL, VPN, VoIP, virus scanning, etc.
- **Key IP: packet scanning, classification, & ultra-fast dispatch to threads.**
  - **RTOS, privileged mode code.**
- Why? One reason:
  40% of TCP packets are 40-byte TCP ACKs.
  - At 10Gb/s, one every 32 ns.

Vendors: Cavium, P.A. Semi, Raza, Broadcom (SyByte), PMC Sierra, Freescale

Industry Initiatives for Accelerator Attachment

- **Geneseo**: Extending PCI Express® (ann. 9/26/06)
  - Proposed by Intel and IBM
    - 30+ companies support at announcement
  - Goal: new open standard for attaching accelerators and co-processors to server platforms
    - power management, transaction ordering (on PCIe), atomic ops, software overhead, faster data rate
  - Developing through and with the PCI SIG
    - May ultimately be called “PCI 3.0” or “2.1” or the like.
- **Torrenza**: Extending HyperTransport™ (ann, 6/1/06)
  - Proposed by AMD
    - 6+ companies support at announce, plus consortia
  - Goal: improve support for the integration of specialized coprocessors in systems based on AMD Opteron microprocessors
    - both I/O (HT) and SMP socket (cHT).
  - Allied with HyperTransport Consortium and OpenFPGA consortium

The Rest of My Answers

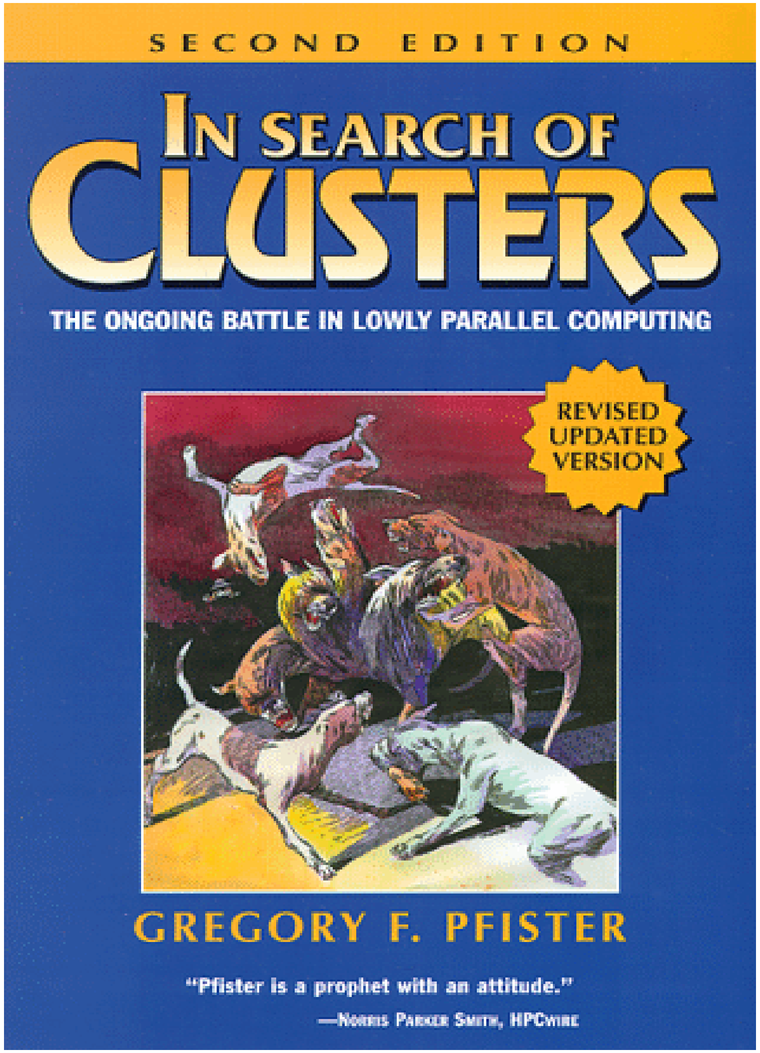
Will full or partial offload be the norm for Ethernet?	Partial
Will any collective functions *typically* be offloaded to the NIC in the near future?	No
What about collective support in switches **{*typically* }**?	Heck no
Should NICs assist in "read-modify-write" capability to support remote updates?	No, but.

Required

What communication functions should **always** be offloaded to accelerators?	ACK ⇒ RDMA, packet classification, work “dispatch”
Which should **never** be offloaded to accelerators?	Everything else.
Are there cases for which it makes sense to offload to another core rather than to an accelerator? (For SMT cores, I suppose one might also consider offloading to other threads of the same core?)	That’s where the “everything else” goes.

...

- Thank you for listening.
- Any (more) questions?


SECOND EDITION
IN SEARCH OF CLUSTERS
THE ONGOING BATTLE IN LOWLY PARALLEL COMPUTING
REVISED UPDATED VERSION
GREGORY F. PFISTER
"Pfister is a prophet with an attitude."
—Norris Parker Smith, HPCwire

Just in case any of you were wondering...

(No, I can't give a presentation without plugging my book.)

Extremely nonrandom clipart