Server Architectures: Input/Output

January 2005

René J. Chevance

Foreword

- **This presentation is an introduction to a set of presentations about server architectures.**
  They are based on the following book:

  Serveurs Architectures: Multiprocessors, Clusters, Parallel Systems, Web Servers, Storage Solutions
  René J. Chevance
  Digital Press December 2004 ISBN 1-55558-333-4
  http://books.elsevier.com/

  This book has been derived from the following one:

  Serveurs multiprocesseurs, clusters et architectures parallèles
  René J. Chevance
  Eyrolles Avril 2000 ISBN 2-212-09114-1
  http://www.eyrolles.com/

  The English version integrates a lot of updates as well as a new chapter on Storage Solutions.

Contact: www.chevance.com **rjc@chevance.com**

Organization of the Presentations

- Introduction
- Processors and Memories
- ➔ Input/Output (this presentation)
  - Architecture
  - PCI
  - SCSI and ATA
  - Fibre Channel
  - Virtual Interface Architecture
  - *Note: Data storage is covered in a specific chapter*
  - Networks and Communications Subsystems
- Evolution of Software Technologies
- Symmetric Multi-Processors
- Cluster and Massively Parallel Machines
- Data Storage
- System Performance and Estimation Techniques
- DBMS and Server Architectures
- High Availability Systems
- Selection Criteria and Total Cost of Possession
- Conclusion and Prospects

Few Comments

- **I/O is an important dimension of system architecture**

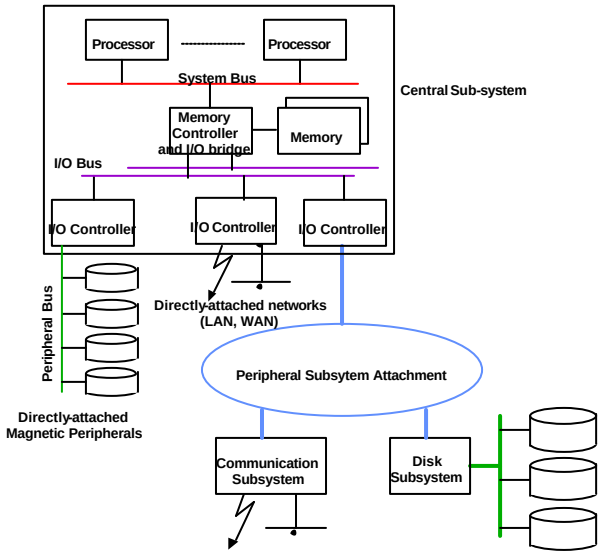
but

- **Until a recent past, most of the system architects and researchers have not paid too much attention to I/O**

- **I/O is an area where it is relatively more easy to introduce distinctive differentiation (as compared with Instruction Set Architectures for instance)**

Generic Architecture of an I/O System

- **Elements of an I/O system**



Evolution of I/O

- **Industry has moved from proprietary I/O architectures towards the use of standards:**
  - **PCI**
  - **ATA, SCSI**
  - **Fibre Channel**
- **Industry initiatives to promote new standards and to develop corresponding technologies (e.g. chip sets)**
  - **I_2O (Intelligent I/O)**
  - **VIA (Virtual Interface Architecture)**
  - **InfiniBand**
- **Emerging concept of Storage Area Network (SAN)**

PCI

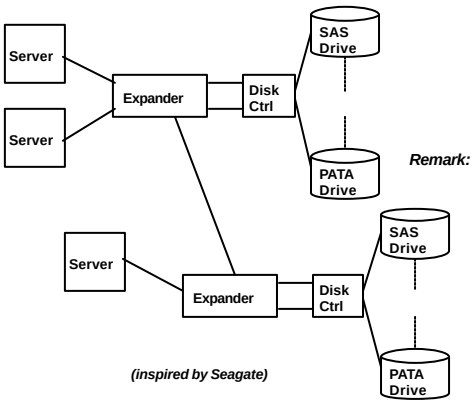
- Standard originating from the PC industry
- Characteristics:
  - Two data transfer widths - 32 bits and 64 bits
  - Two clock rates - 33MHz and 66MHz
  - Various available bandwidths (133, 266 and 532 MB/sec)
  - Simultaneous support of 32 and 64 bits controllers on the same bus
  - Plug and play (automatic discovery of the configuration)
  - Hot plug
- Evolution towards PCI-X (but maintaining ccompatibility) with bandwidth up to 4256 MB/ sec (PCI-X 2.0)

SCSI and ATA

- SCSI has long been a de facto standard for peripheral connection.
  - Limitations: fairly limited maximum length for a connection, and bulkiness and weight of the connectors and cables
    - Up to 15 devices
    - 40, 80, 160 or 320 MB/s
    - up to 25m (~80 ft.)
    - 68-wire cable
  - Evolution towards SAS (Serial Attached SCSI)
    - 3 GB/s (12 GB/s planned)
    - dual ported drives
    - point to point connections
    - extended drive addressability (up to 16,286 devices per port) and connectivity
    - capability to connect SATA drives (SATA signals are a subset of SAS signals).
- Emergence of ATA (Advanced Technology Attachment) coming from the PC industry
  - PATA : flat cable 40 wires, 100 MB/s
  - SATA (Serial ATA) : serial interface 150 MB/s
  - SATA II (2004): 300 MB/s

SCSI and ATA(2)

- **Exploiting properties of SAS**



Remark: Selection of SAS drives for performance and SATA drives for cost

(inspired by Seagate)

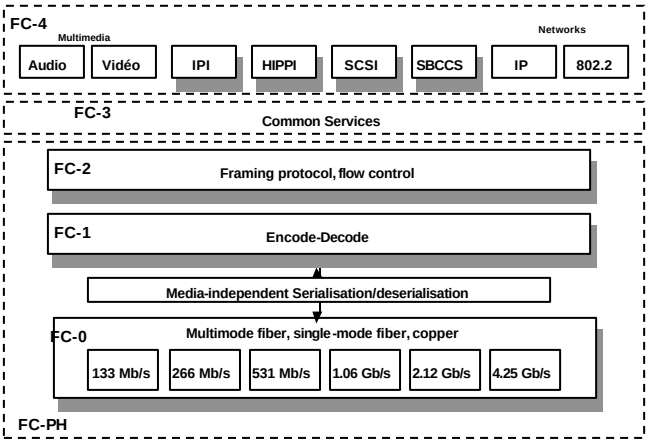
Fibre Channel

- **Fibre Channel: serial interface:**
  - **Fibre Channel Arbitrated Loop (FC-AL) for peripheral connection**
  - **Fibre Channel for connection to sub-systems and/or inter-system connection**
  - **Fibre Channel allows the support of different protocols**
- **Comparison of peripheral interfaces, networks and Fibre Channel**

I/O Channels	Networks	Fibre Channel
• Normally implemented in hardware	• Normally imple-mented in software	• Implemented in hardware
• High speed - O(10 MB/s)	• Moderate speed - O(MB/s)	• High speed - O(100MB/s)
• Limited connectivity	• Rich connectivity	• Rich connectivity
• Low latency - O(µs)	• Moderate latency- O(100 µs)	• Low latency
• Short distance	• Long distance	• Long distance (with optical)
• Master/slave archi-tecture	• Peer-to-peer archi-tecture	• Both master/slave and peer-to-peer supported
• Strong data integrity	• Fragile	• Strong data integrity
• Error detection at very low level (hardware)	• Error detection at a very high level (in software)	• No management station

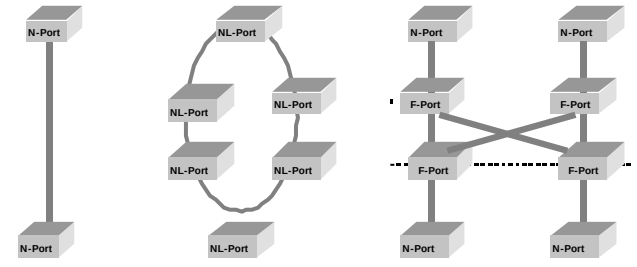
Fibre Channel(2)

- **Layered Architecture of Fibre Channel**



Fibre Channel(3)

- **Fibre Channel Topologies**



Point-to-Point
100 MB/sec

FC-AL Loop
100MB/sec

Fabric
N x 100 MB/sec

Fibre Channel(4)

Comparison of SCSI and Fibre Channel

Property	SCSI	Fibre Channel (Loop)
Number of units per phys-ical connection	15	126
Bandwidth	40, 80, 160 or 320 MB/s	100MB/s
Length of a physical con-nection	up to 25m (~80 ft.)	Copper: 30m (~100ft) 5µm or 50µm Fibre : 175-500m (~500-1600 ft.) 9µm Fibre: several kilometers (miles)
Nature of the physical interface	68-wire cable	4-wire cable or 2-Fibre optical interconnect

Fibre Channel Service Classes

	Class 1	Class 2	Class 3	Class 4
Service Characteristics	• dedicated - the whole of the bandwidth is available	• multiplexed, datagram service	• multiplexed, datagram service	• a fraction of the bandwidth is dedicated to transport of realtime data, virtual circuits with quality of service guarantee for mini-mum bandwidth and bounded latency
	• fixed routing	• adaptive routing		• congestion-free
	• in-order message delivery	• in-order delivery of packets not guaranteed	• no confirmation of delivery or of failure to deliver	• in-order packet delivery guaranteed
	• confirmation of delivery or of failure to deliver	• confirmation of delivery or of failure to deliver		
	• end-to-end flow control	• end-to-end flow control and per-connection flow control	• flow control available solely at the level of the connection	• access guaranteed
Communication model	• connection-oriented	• connection-less	• connection-less	• connection-oriented
	• circuit-switched	• packet-switched	• packet-switched	• circuit-switched

Page 13

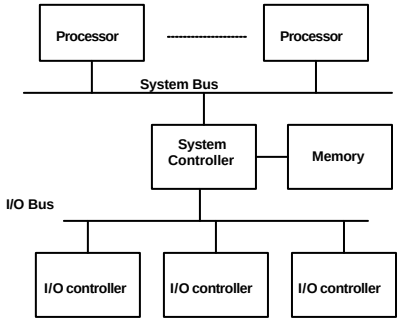
A New I/O Architecture

InfiniBand

Page 14

Problems with « classical » I/O Architecture

- **Several industry initiatives should improve the server I/O (InfiniBand, I_2O (Intelligent I/O) and VIA (Virtual Interface Architecture)**
- **Classical « PC-like » I/O Architecture**



Difficulties:
- **The system controller is a complex design**
- **Limited number (5 or 6) of controllers per I/O bus (multiplication of the I/O buses for large configurations)**
- **Lack of failure isolation**
- **The system controller and the I/O bus can be bottlenecks**
- **A shared I/O bus does not help failure localization**
- **The dialog between the system and the controllers has an impact on performance**
- **With « memory mapped » architecture, a controller may corrupt main memory**

Page 15

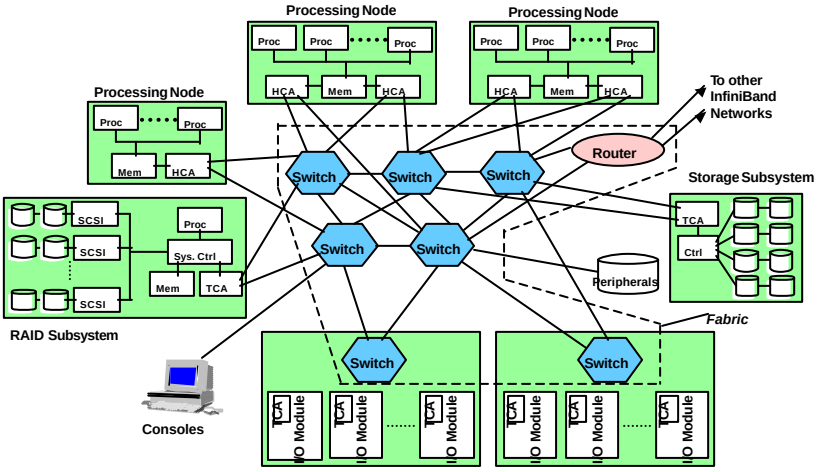
InfiniBand

- **Architecture Objectives:**
  - **Scalability**
  - **Low latency and low overhead**
  - **High Bandwidth (to cope with microprocessor performance evolution)**
    - **256 MB/s up to 6 GB/s**
  - **High Availability**
  - **Unified communication (processes, storage, networking)**
  - **Capabilities of the communication protocol:**
    - **Flow control (static, dynamic, Quality of Service QoS)**
    - **Partitioning**
    - **Multicast**
    - **….**
  - **Low cost due to standardization and availability of components and solutions**

Page 16

InfiniBand(2)

InfiniBand Architecture



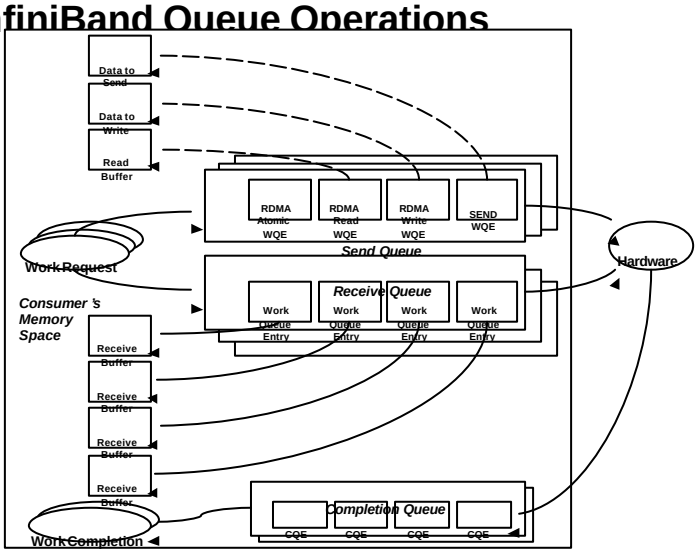
(Source: InfiniBand Trade Association)

Collection of I/O Controllers

Collection of I/O Controllers

InfiniBand(3)

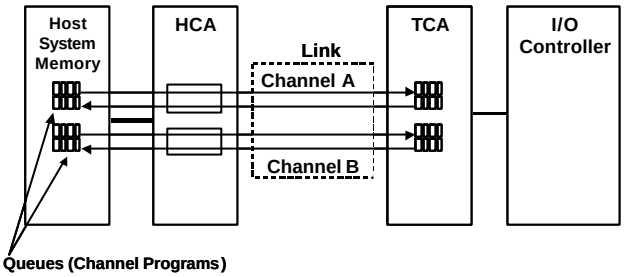
InfiniBand Queue Operations



(Source: InfiniBand Trade Association)

InfiniBand(4)

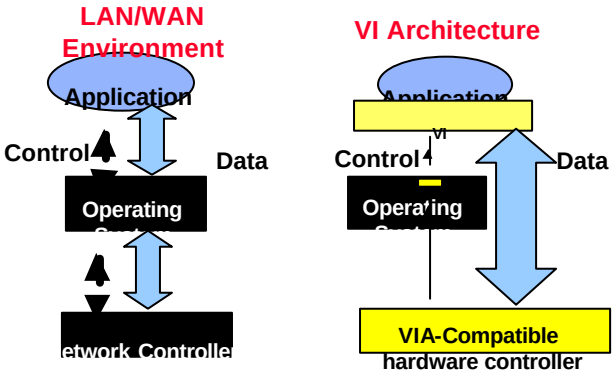
- **Concept of a channel program (inspired by InfiniBand)**
  - Logical connection between two address spaces
  - DMA (Dynamic Memory Access) capability
  - Concept of queues (or channel programs, heritage from mainframe architecture)



Page 19

Virtual Interface Architecture

- **VIA provides a low-latency, high bandwidth data transfers between nodes and storage subsystems in loosely-coupled systems such as clusters**
- **Comparison of VIA and Traditional Communications**



(Source Intel)

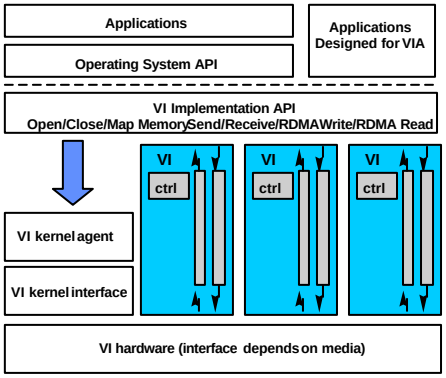
- The operating system is involved only in the establishment and tear-down of the connection and in error handling

Page 20

Virtual Interface Architecture(2)

Elements of VIA

- VIA s a form of RDMA (Remote DMA)



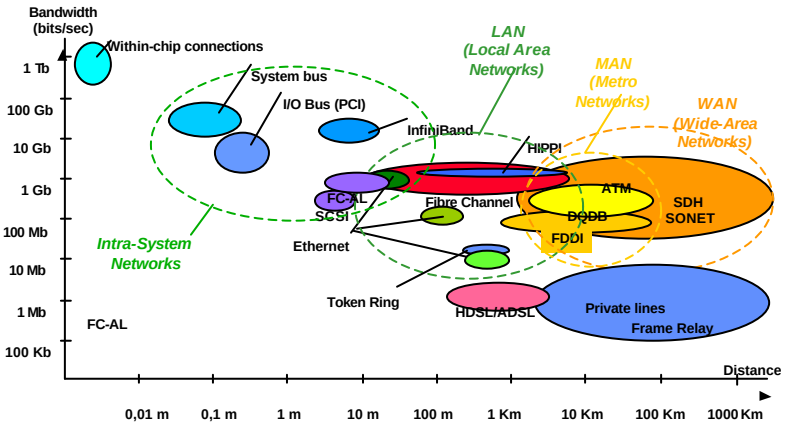
(Source Intel)

Elements of VIA

- Send and receive packet descriptors (with *scatter-gather* operations)
- A send message queue and a receive message queue (comprising linked lists of packet descriptors)
- A means of notifying the network interface that packets have been placed on a queue
- An asynchronous notification process for the status of the operations requested
- Registration of memory areas used for communications

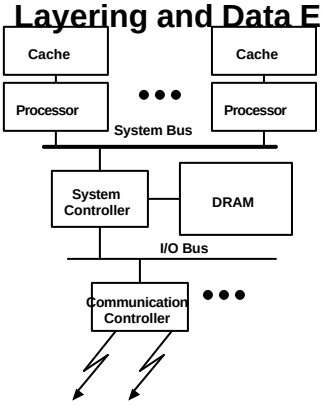
Networks and Communications Subsystems

Communication Technologies

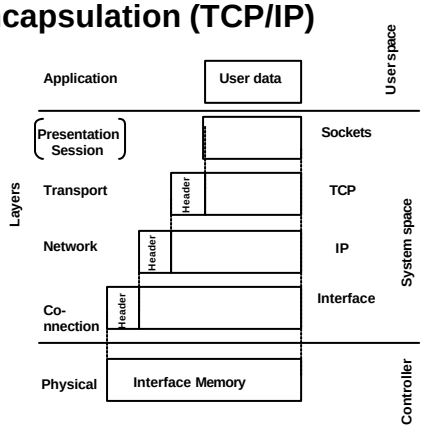


Networks and Communications Subsystems(2)

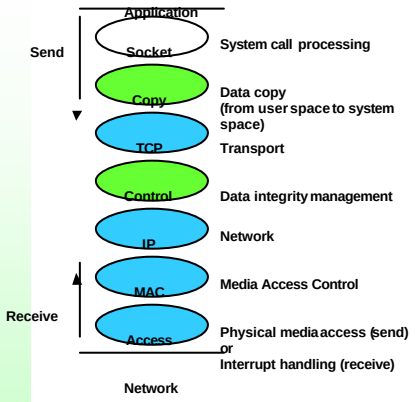
- **Server Support for Communications (inspired by [MCK99])**
  - **Problems: Very large number of interrupts, Memory Copying, Cache Interactions**
- **Layering and Data Encapsulation (TCP/IP)**



«Classical » Architecture



Networks and Communications Subsystems(3)



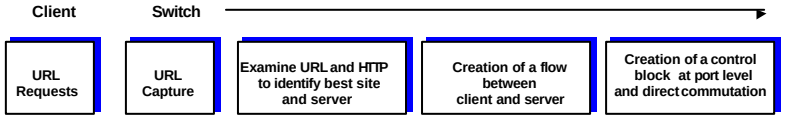
Logical Processing Steps in TCP/IP

Legend:
- *White indicates per-message processing*
- *Blue indicates per-fragment processing*
- *Green indicates per-byte processing*

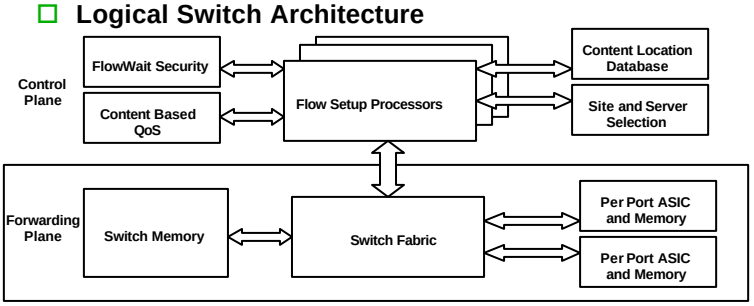
- **Optimizations**
  - **Reducing the number of interrupts:**
    - **DMA with linked lists of descriptors**
  - **Reducing data movements**
    - **DMA with linked lists of descriptors**
    - **Specialized chip for checksum computation**
  - **The above elements are reducing processor involvement and so, cache intreractions**

New Generation of Switches

- **Web Switching (Level 5 TCP/IP) -WebSwitch (ArrowPoint - CISCO)**
  - **Logical processing steps**



  - **Logical Switch Architecture**

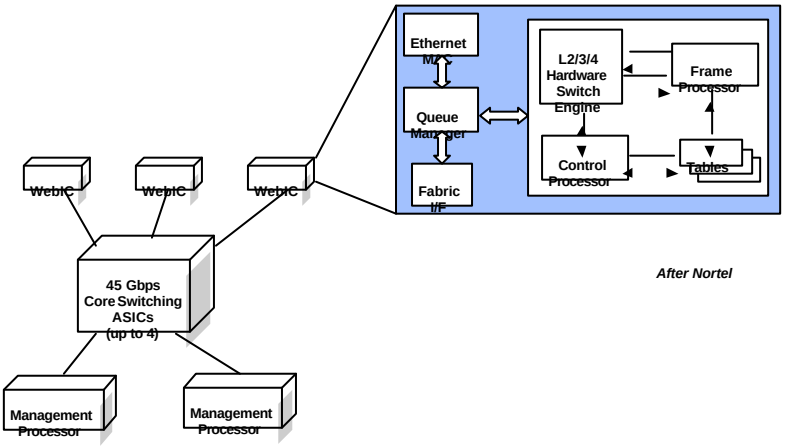


After ArrowPoint

Page 25

New Generation of Switches(2)

- **Nortel**



After Nortel

Page 26

References

[MCK99] **Nick McKeown, «An Overview of Techniques for Network Interface Design»,**
Course EE384c, High Performance Network Systems,
Stanford University, *http://www.cs.stanford.edu*.