Minimizing PCI Express® Power Consumption

Akber Kazmi

Director of Marketing

PLX Technology

Agenda

- Why power consumption is important?
- Major contributors
- Power consumption of PCI Express® components
- Considerations for PCI Express ASIC/ASSP design
- Considerations for system/board design with PCI Express
- Summary

Why Power Is Important?

- A 2006 survey* found:
  - ✓ Electricity use of servers doubled between 2000 and 2005
  - ✓ In 2005, electricity bill for server users was $2.5B in U.S. and $7.3B worldwide
  - ✓ In 2005, server farms used 1.2% of total U.S. electricity consumptions
  - ✓ Average electricity bill for a 300-watt computer would be around $450/year
    - Varies with power rates, power saving options, etc.
  - ✓ Consumer, enterprise and environment need power efficient computer systems

*** J. Koomey study**

Tip of the Iceberg

- Servers: a fraction of computers shipped and in use (2005 data)
  - ✓ 7M Servers shipped
  - ✓ 27M Servers in use
  - ✓ Electricity bill (servers): $7.3B
  - ✓ 197M PCs shipped in 2005
  - ✓ One Billion PCs in use
  - ✓ Electricity bill (PCs): $35B*

***Assume: no cooling cost & 6 hours of use**

Power Hungry

- Demand for bandwidth continues to grow
  - ✓ Bandwidth = power
- By 2010, expected growth in overall power use by servers to be 76% over 2005
  - ✓ Expect similar growth in enthusiast desk-top systems
- Cost of power will exceed hardware cost in three years*
  - ✓ With only 30% growth in bandwidth
  - ✓ No change in performance/watt
- Need power efficiency and awareness

*** "The Price of Performance" by Luiz Barrosa**

Cost of a Watt

Item	Units	Comments
Device Typical Power Consumption	100 W	Example
Actual Power Consumed (power supply efficacy of 80%)	125 W	Varies from 30% to 90%
Cost per hour = Cost/KWH * power Consumption/1000 => 7x125/1000 cents	$0.0875	Range: 4-16 cents/KWH, Example uses: $0.07
Cost/week (24x7 operation)	$1.47	
Cost/year (52 weeks)	$76.44	
Cooling cost multiplier (x2)	$152.88	Amortize fans and cooling equipment electricity bill
Cost Per watt/year	$1.5288	

Major Contributors

- Servers, workstations and desktops
- Key components
  - ✓ Processor – 50-100 W
  - ✓ Chipset – 10-30 W
  - ✓ Switches/bridges – 5-8 W
  - ✓ Graphics cards or GPUs– 50-200W
  - ✓ I/O endpoints (disk, sound, DVD) – 20-50 W
  - ✓ Memory – 10 W
  - ✓ Monitor – 70-150W
  - ✓ Power supply loss – 20-50W (10-30% loss)

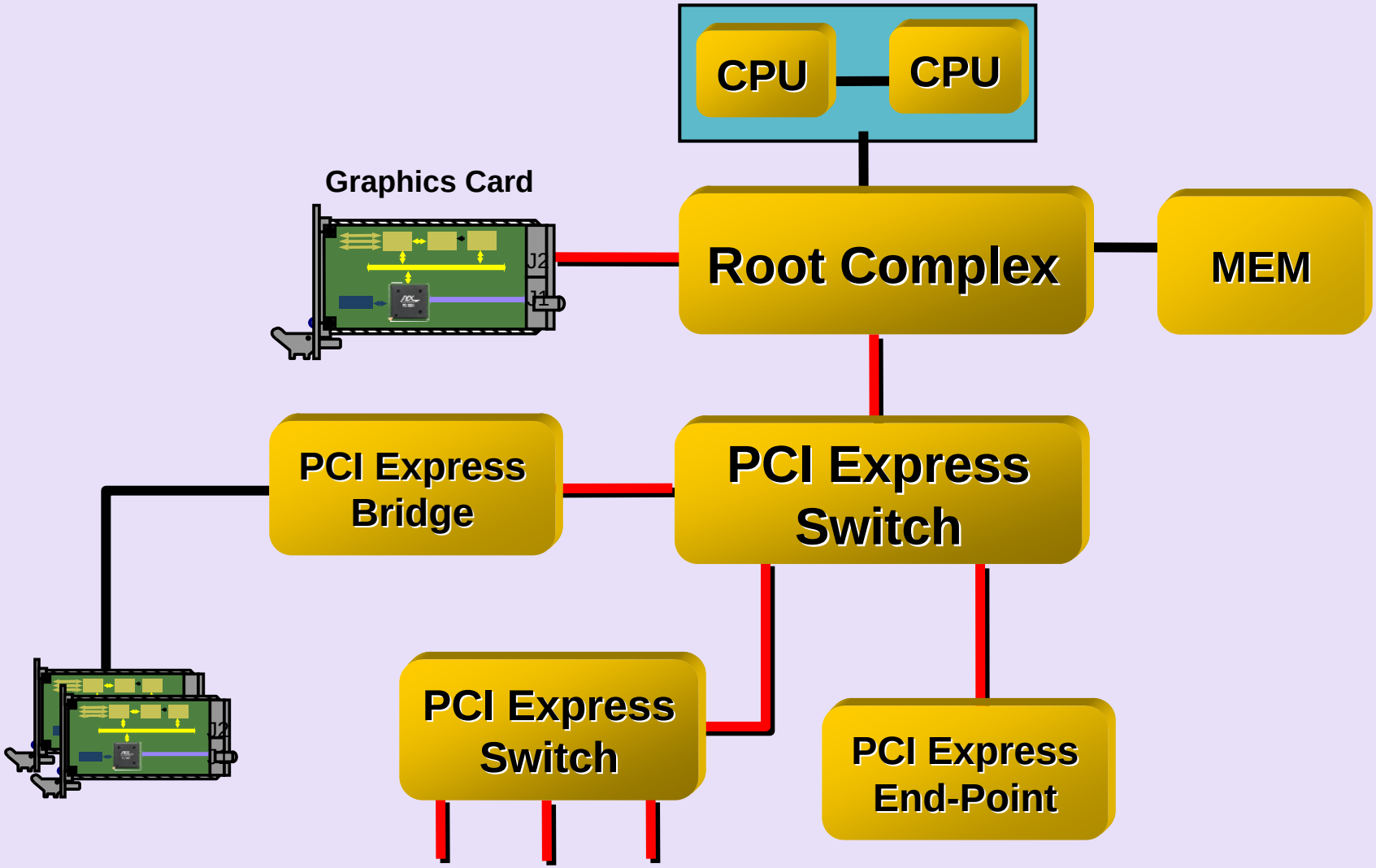
Industry & Gov. Efforts

- ENERGY STAR power management
  - ✓ EPA initiative to help power management in Windows/Mac PCs
- The Green Grid
  - ✓ A global consortium developing and promoting energy efficiency for data centers and information service delivery
- OnNow
  - ✓ A design initiative to ensure new PCs are instantly accessible to users and take advantage of the vast improvements in PC power management technology
- Common Power Format
  - ✓ A power-saving technique in silicon design
- Many other initiatives underway

PCI Express Power Management

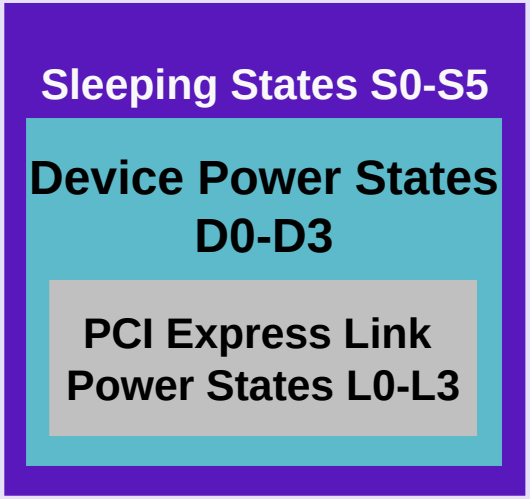
- PCI Express is a key interconnect standard
  - ✓ Currently runs at 2.5Gbps (5Gbps being tested; 10Gbps on the horizon)
  - ✓ Chipset and IO devices connect through PCI Express bridges and switches
  - ✓ High-speed SerDes contribute heavily in switch power consumption
- Advance power management features shall be implemented and used

PCI Express Components



PCI Express Power Management

- Builds on PCI power management (PM)
  - ✓ Compatible with PCI PM software stacks
- Three levels
  - ✓ System level,
  - ✓ Device level
  - ✓ Link level



Additional Power Management

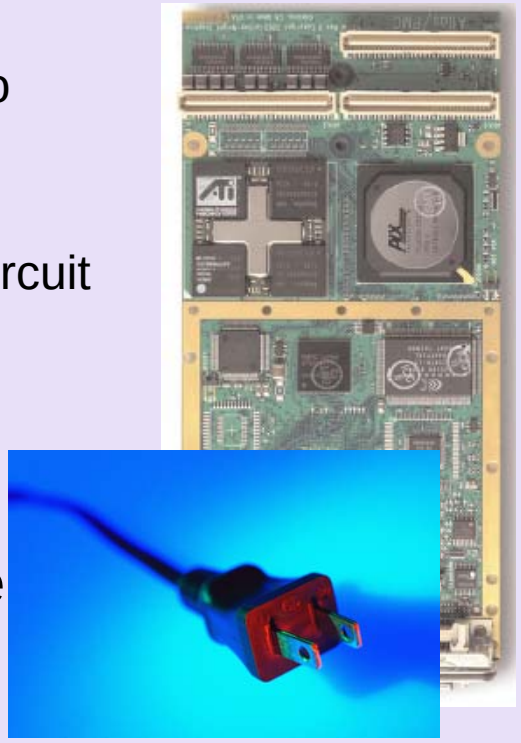
- Enhanced PM capabilities
  - ✓ Power reduction through active state PM (L0s, L1)
  - ✓ PME using in-band messaging
  - ✓ SW control of Vaux
- Minimize Flow Control overhead to maximize bandwidth
  - ✓ Single FC update and acknowledge for multiple packets
  - ✓ Recommend 3-4 transactions per FC update/Ack

Power Management Event (PME)

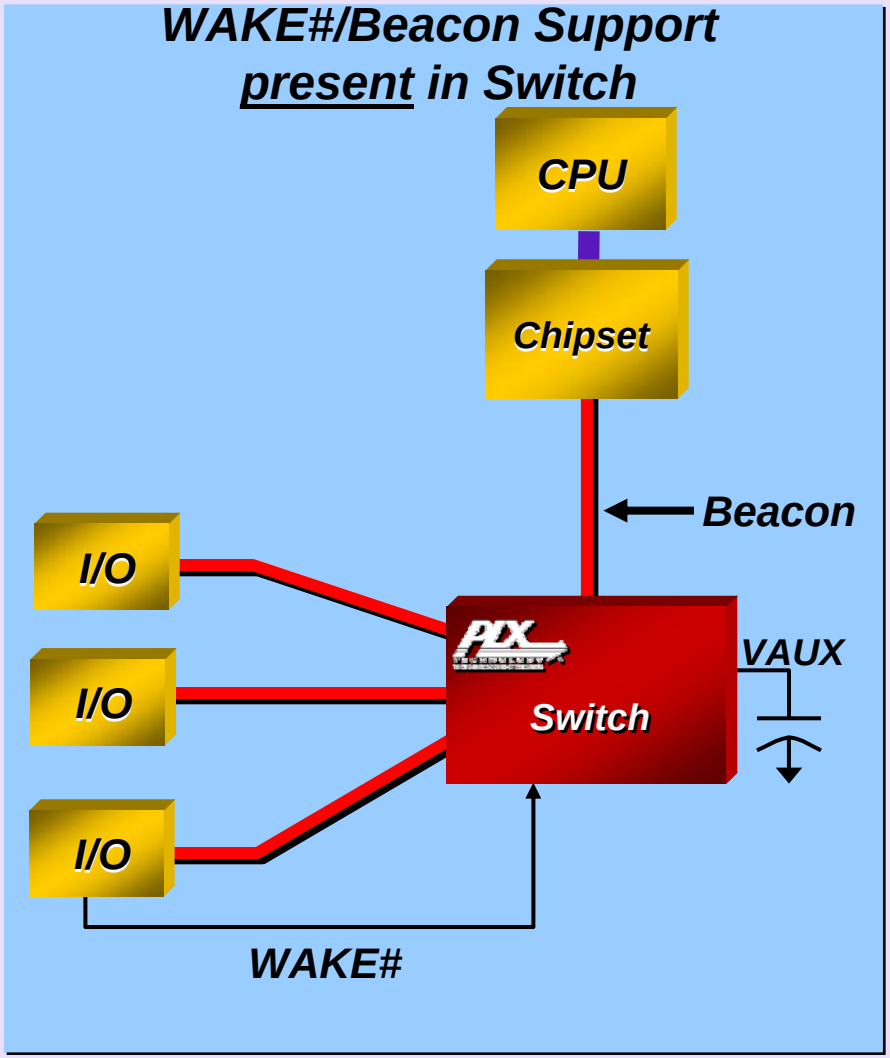
- Wake event management
  - ✓ Wakeup using WAKE# or Beacon
  - ✓ PME Message following wakeup
- Active State Power Management (ASPM) provides *additional* benefit
  - ✓ Low latency – minimum impact on performance
  - ✓ Finer granularity of control – more opportunity for power savings compared to software controlled PM
- Serial signaling technology consumes power when not sending data
  - ✓ PCI Express uses ASPM to reduce power in logical idle

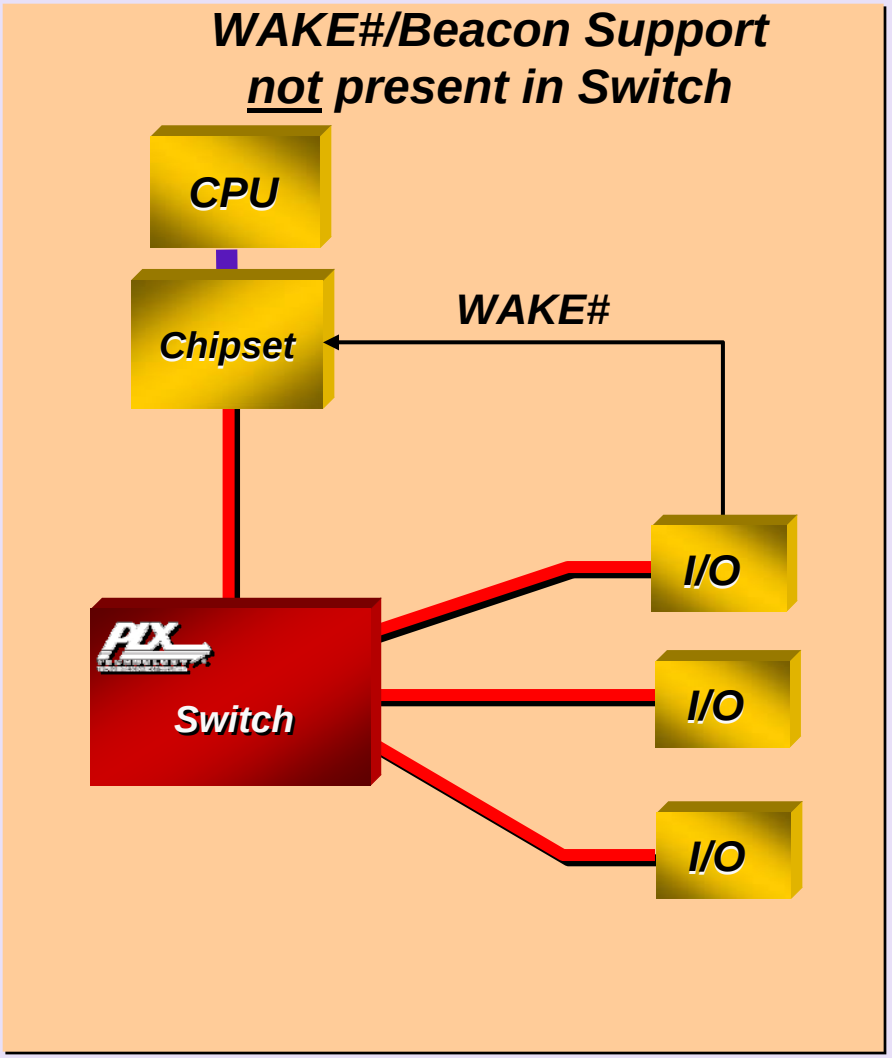
Out of Band Power Management

- Wake/Beacon Management Support
  - ✓ WAKE#
    - – Out of band mechanism used by endpoints to inform host of power state change
  - ✓ Beacon
    - – In-band mechanism used by PCIe® devices to inform host of power state change
  - ✓ VAUX
    - – Auxiliary voltage supply for Beacon internal circuit
- VAUX/WAKE#/Beacon support
  - ✓ WAKE# - Input Signal to Switch
  - ✓ Switch generates in-band Beacon seque to host when WAKE# is active

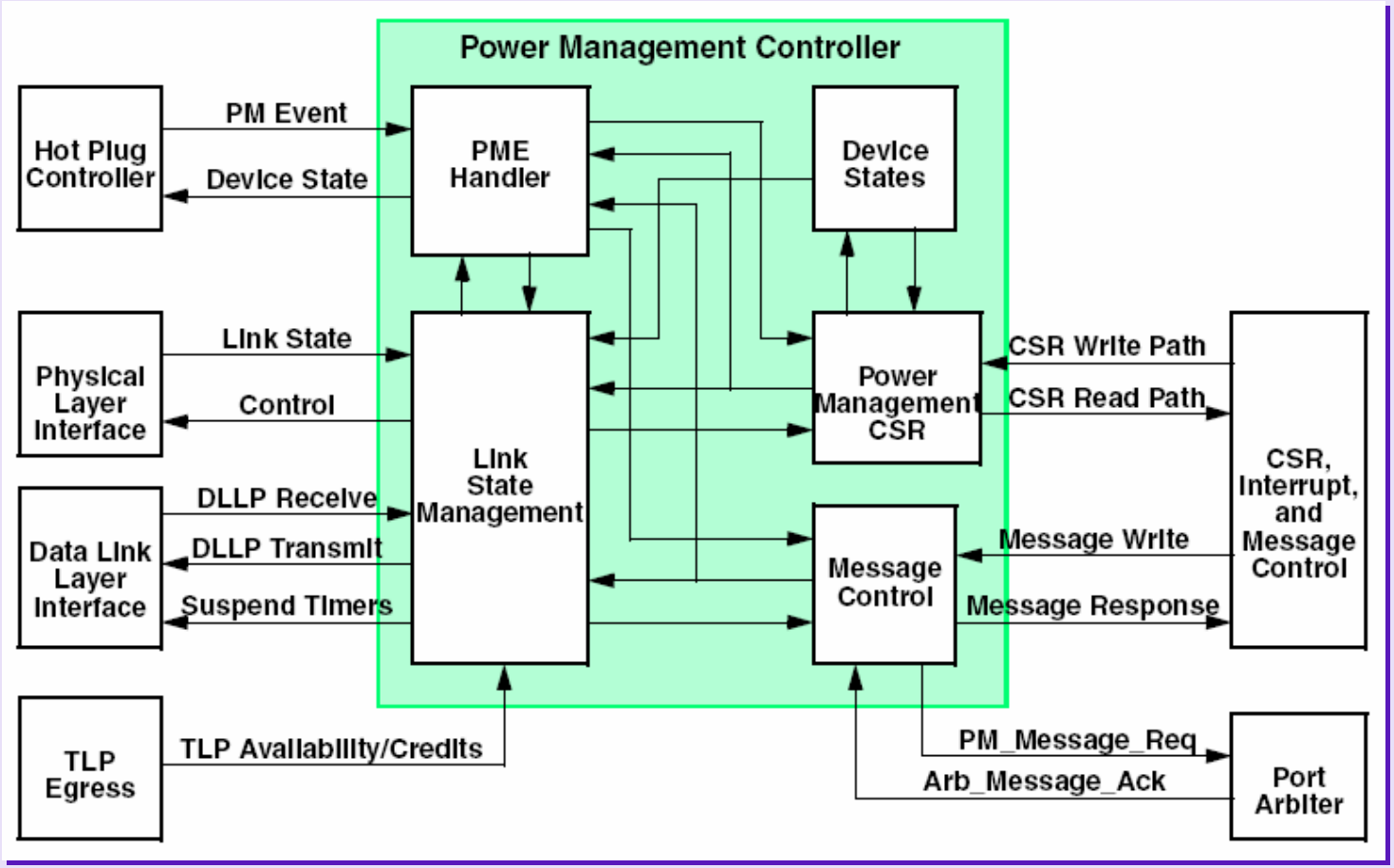


VAUX/WAKE#/Beacon





Power Management Controller



Typical PM Controls

- Hardware autonomous
- Software driven
- Link States (L-states)
  - ✓ Hardware driven states
- Devices States
  - ✓ D0_uninitialized after power-up
  - ✓ D0_active after device initialized
  - ✓ D3hot and D3cold: software controlled
- Events: Hot-plug, D3hot

Power Consumption of ASIC/ASSP

- Power consumption of a device depends on:
  - ✓ Number switching gates
  - ✓ SerDes speed
  - ✓ SerDes voltage
  - ✓ Core voltage
  - ✓ Operating temperature
  - ✓ Process technology
  - ✓ Traffic type
  - ✓ PCIe power management support

Considerations for ASIC/ASSP Design

- Plan power management in early stage
  - ✓ Use state-of-the-art design tools
- SerDes selection and optimization
  - ✓ Control drive current
  - ✓ Low leakage
  - ✓ Control over 2.5GT/s and 5GT/s speeds
- Phy Design
  - ✓ Dynamic Link width control
- Optimize design
  - ✓ Low gate count
  - ✓ Clock disable for unused blocks
  - ✓ Control leakage

Considerations for System Design

- Select devices that:
  - ✓ Consumes less power for a given functionality
  - ✓ Support PCIe power management features
  - ✓ Allow SerDes to train up/down dynamically
  - ✓ Allow ports to train up/down dynamically
  - ✓ Allow SerDes to be turned-off when not in use
  - ✓ SerDes drive current controlled to reduce power
  - ✓ Allow part of the core logic to turn-off when not active
- Implement power management features through entire data path

What to Expect?

- 2.5GT/s Switches
  - ✓ 100-150 mW per lane for low lane count
  - ✓ <100 mW per lane for high lane count
- Bridges
  - ✓ Single lane <0.5 W
  - ✓ Four lane ~2 W
- 5GT/s Switches: 150-200 mW/lane

PLX Support for Power Management

PLX® Technology: ExpressLane™ PCI Express SWITCHES						
Part Number	Lanes	Ports	Latency (ns)	Package Size (mm)	Power (W) Typ.	Power (W) Max.
PEX 8505	5	5	138	15 x 15	0.8	1.4
PEX 8508	8	5	150	19 x 19	1.6	2.1
PEX 8509	8	8	118	15 x 15	1.2	1.8
PEX 8516	16	4	275	27 x 27	3.2	4.3
PEX 8517	16	5	150	27 x 27	2.6	3.6
PEX 8518	16	5	150	23 x 23	2.6	3.6
PEX 8524	24	6	275	31 x 31	3.9	6.1
PEX 8525	24	5	115	31 x 31	2.6	3.7
PEX 8532	32	8	275	35 x 35	5.7	7.4
PEX 8533	32	6	115	35 x 35	3.3	4.8
PEX 8547	48	3	110	37.5 x 37.5	4.9	7.1
PEX 8548	48	9	110	37.5 x 37.5	4.9	7.1

PLX® Technology: ExpressLane™ PCI Express BRIDGES			
Part Number	Lanes	Package Size (mm)	Description
PEX 8111	1	10 x 10 / 13 x 13	PCIe to PCI
PEX 8112	1	10 x 10 / 13 x 13	PCIe to PCI
PEX 8114	4	17 x 17	PCIe to PCI-X
PEX 8311	1	21 x 21	Local Bus to PCIe

Summary

- PCs and servers are consuming enormous amount of energy and costing large sums to own
- Industry and governments are supporting several efforts
- Use PCI Express technology that offers excellent features for power optimizations
- Use components that support power management and optimization
- Remember: Every watt counts when running 24x7!

Thank you for attending the
PCI-SIG Developers Conference 2007.

For more information please go to
www.pcisig.com

Minimizing PCI Express Power Consumption

Akber Kazmi

Director of Marketing

PLX Technology