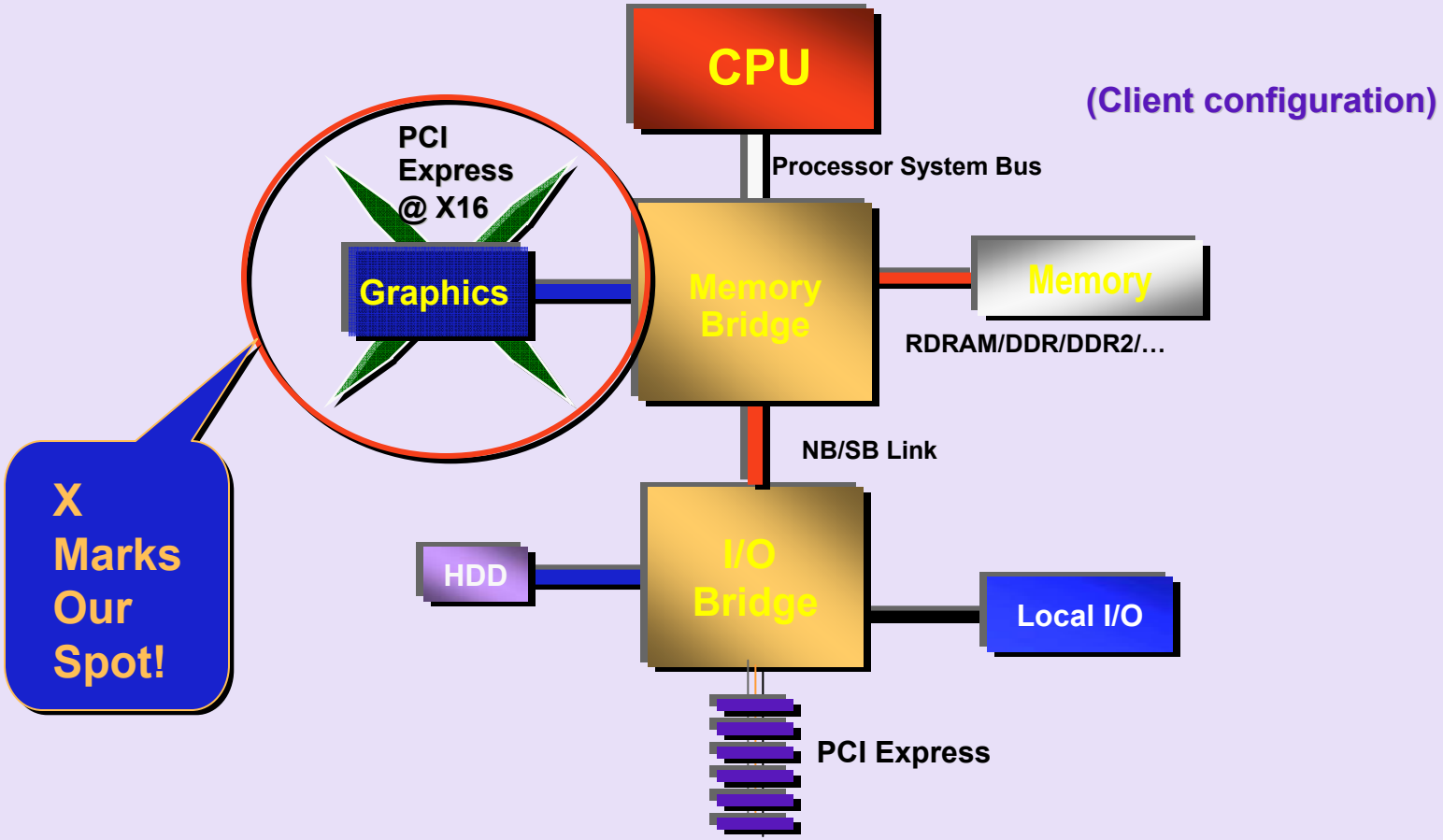
PCI Express Graphics

Presented by Chuck Stancil
Hewlett-Packard Company
Chairman, PCI Express Electromechanical WG

Content provided by Dave Zenz
Dell Inc.
Chairman, PCI Express Graphics WG

PCI Express Platform Implementation Overview



Agenda – PCI Express Graphics

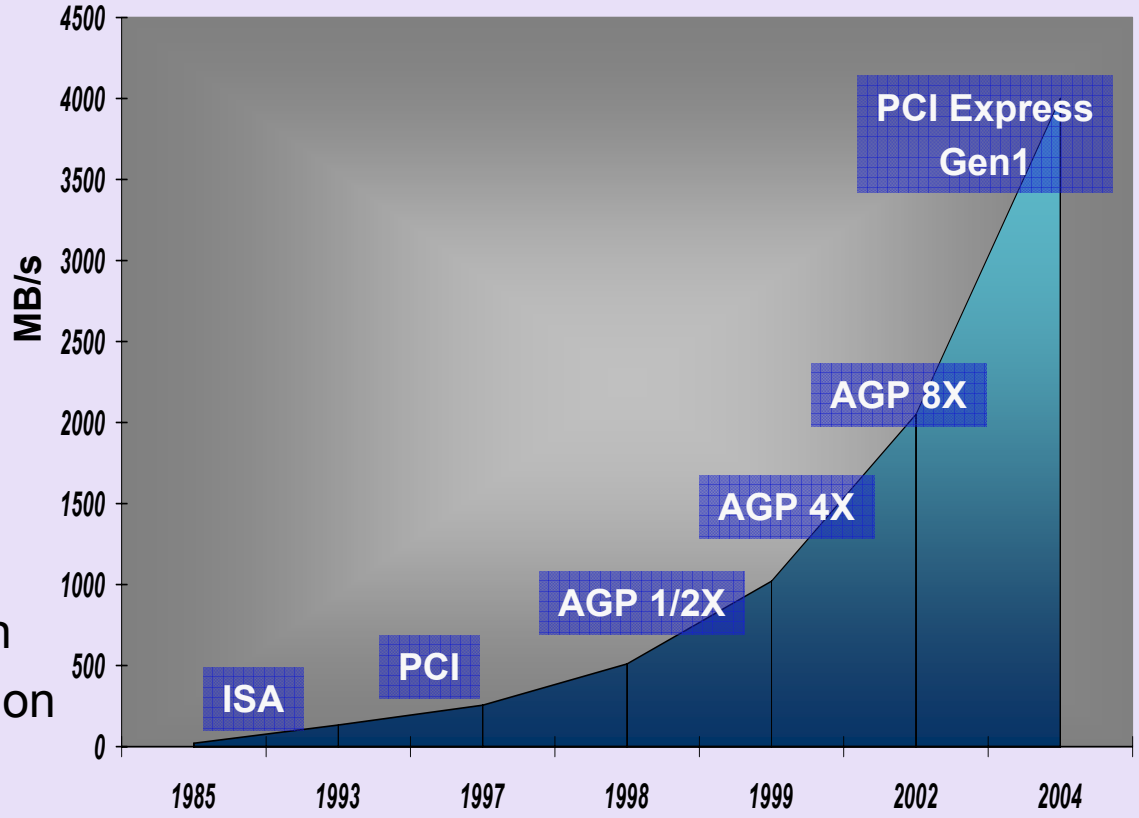
- Introduction to Graphics Interconnects
- PCI Express Graphics Implementation
- Transition strategy
- Summary & Call to Action

Introduction to Graphics Interconnects

- Evolution of Graphics Interconnect
- PCI Express Graphics Goals
- Next Generation Requirements
- Graphics Interconnect and Performance

Evolution of PC Graphics Interconnects

- Early 80's -- ISA
  - ✓ The baseline
  - ✓ 16bit @ 8.33MHz
- Early 90's -- PCI
  - ✓ Driven by 2D GUI acceleration
  - ✓ 32bit @ 33MHz
- Mid/late 90's -- AGP
  - ✓ Driven by 3D HW/SW
  - ✓ 32bit @ 66MHz (1X)
  - ✓ 1X->2X->4X->8X evolution
  - ✓ 3.3V -> 1.5V -> .8V evolution
- 2004 – PCI Express: The Next Generation



Goals for PCI Express Graphics

- Significant differentiation over AGP
- Provide clear scaling targets on future roadmaps
- Comprehend all segment requirements
- Transparent transition for end user

Next Generation Requirements

- Bandwidth *roadmap*
  - ✓ Beyond one last gasp on AGP
- Enable emerging and future applications
  - ✓ E.g. Glitchless streaming media support
- Supports broad range of platforms
  - ✓ Desktop, mobile, workstation, server…
- Friendly to future process technologies
  - ✓ Same old need -- this time a new phy layer opportunity
- Cost Effective Solution

PCI Express Graphics Cost

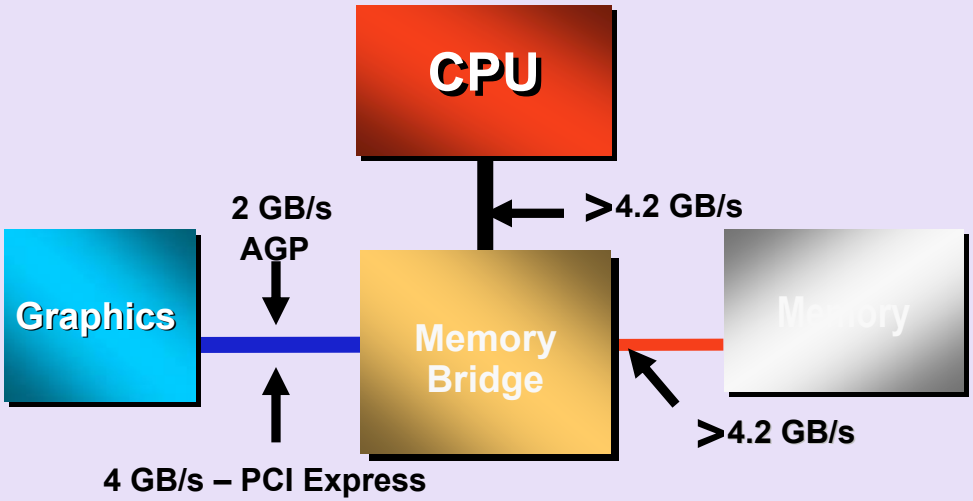
- The BW/pin win with PCI Express can improve costs in numerous areas
  - ✓ 133MB/s PCI vs. X1 PCI Express @250MB/s
- Silicon area requirements are similar to AGP for ’04 process technologies, and improve over time
- X16 PCI Express connector in volume to be at cost parity to AGP’s
  - ✓ Similar size (89mm)
  - ✓ Simpler construction

Higher Performance With Cost Equivalence

Graphics Interconnect & System Balance

- *Maintaining balanced system bandwidth is critical*

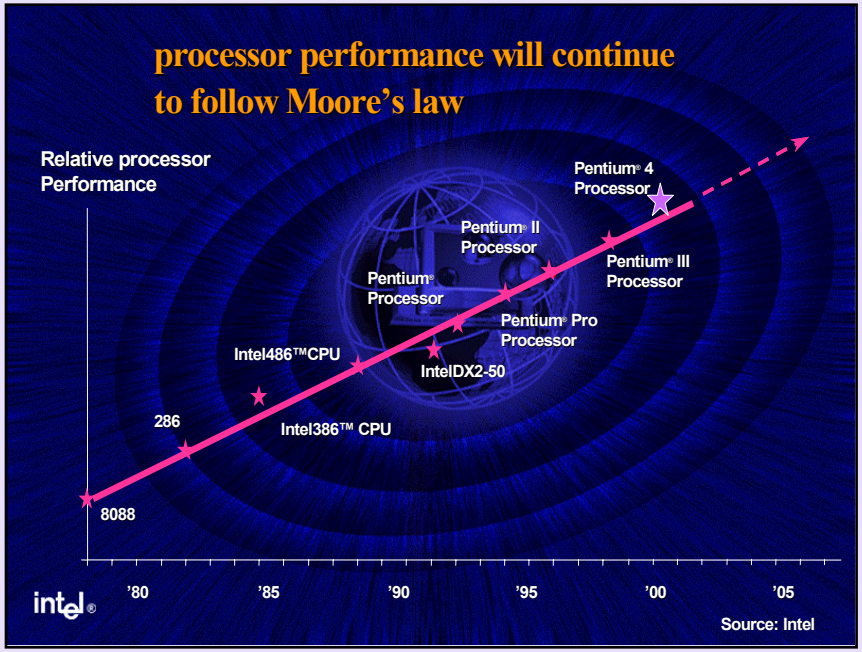
- As CPU ↔ Memory feeds & speeds continue to grow
  - ✓ So too must the graphics interconnect bandwidth

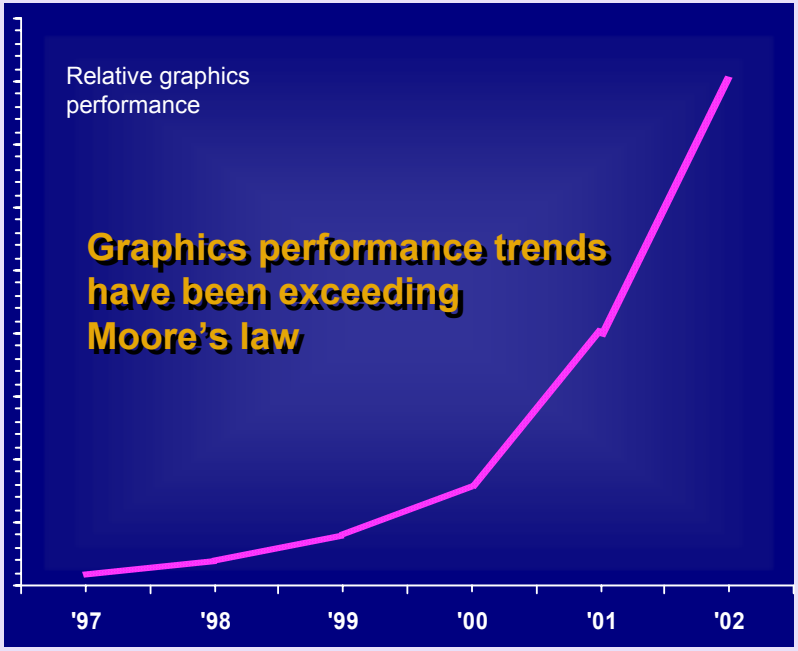


PCI Express Will Provide the System Balance Required for Graphics

Graphics Interconnect & System Performance

- *Maintaining and enabling access to all components of system performance, in our case graphics, is critical*
- As CPU performance continues to escalate
  - ✓ So too does graphics performance, but even at a faster rate

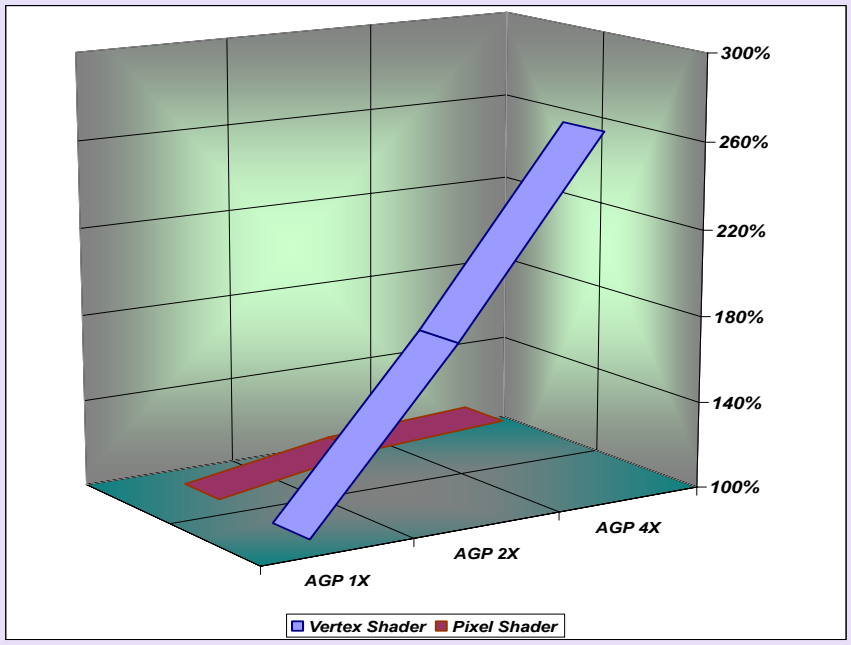


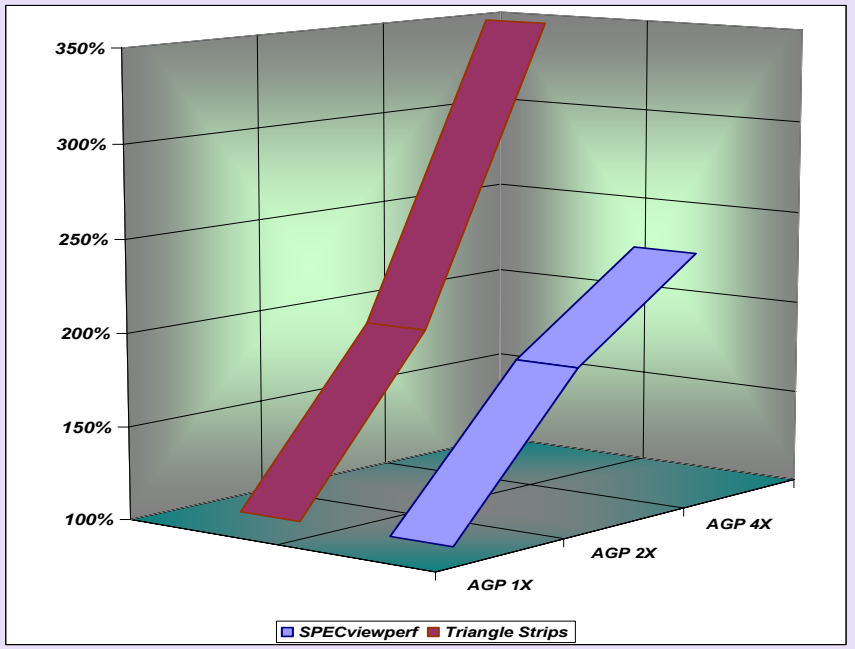


PCI Express Will Enable Continued Access to the Increasing Power of Graphics

Graphics Benchmark Data – Consumer & Workstation

- Graphics performance can be limited by today's graphics interconnect
  - ✓ Scaling with AGP speed indicates there's untapped graphics performance available





Graphics Interconnect & Software Optimizations

- Today's software, be it a driver, an application or an OS, are tuned to utilize the available graphics bandwidth

- Software optimizations to take advantage of additional graphics bandwidth will come
  - ✓ ISVs are telling us that they'll need it
  - ✓ But we must first build it…

Agenda – PCI Express Graphics

- Introduction to Graphics Interconnects
- PCI Express Graphics Implementation
- Transition strategy
- Summary & Call to Action

PCI Express Graphics Implementation

- Bandwidth & Scalability
- New Features & Enhancements
- Shared Memory Model
- Form factor & Power

PCI Express Graphics Bandwidth

- Show me the bandwidth!
  - ✓ Industry expectation is min 2X improvement over existing I/F
- Initial phy layer provides 250MB/s per lane
  - ✓ X16 port target for initial graphics implementations

1st Generation Delivers 8GB/sec Concurrent Peak
– Plus *Scalability*

Scalability

- Scaling, where do we go after the introduction?
- We have flexibility… two choices:
  - ✓ Faster
  - ✓ Wider
- Likely candidate is an evolutionary transition to minimize impact on motherboard layout/size and potentially chassis design
- So Gen 2 will likely go faster

New Graphics Usage Models

- The realization of concurrent real-time graphics applications
  - ✓ Enables the PC to be a true graphics multi-processing device
    - Use graphics video features to process video (PVR functionality) while still using the PC
  - ✓ Allows usage of PC for any application at anytime
- "Real" real-time 3D
  - ✓ Your child can direct their own "Finding Nemo" or climb Mons Olympus
  - ✓ Bring offline rendering online to appease our crave for instant gratification

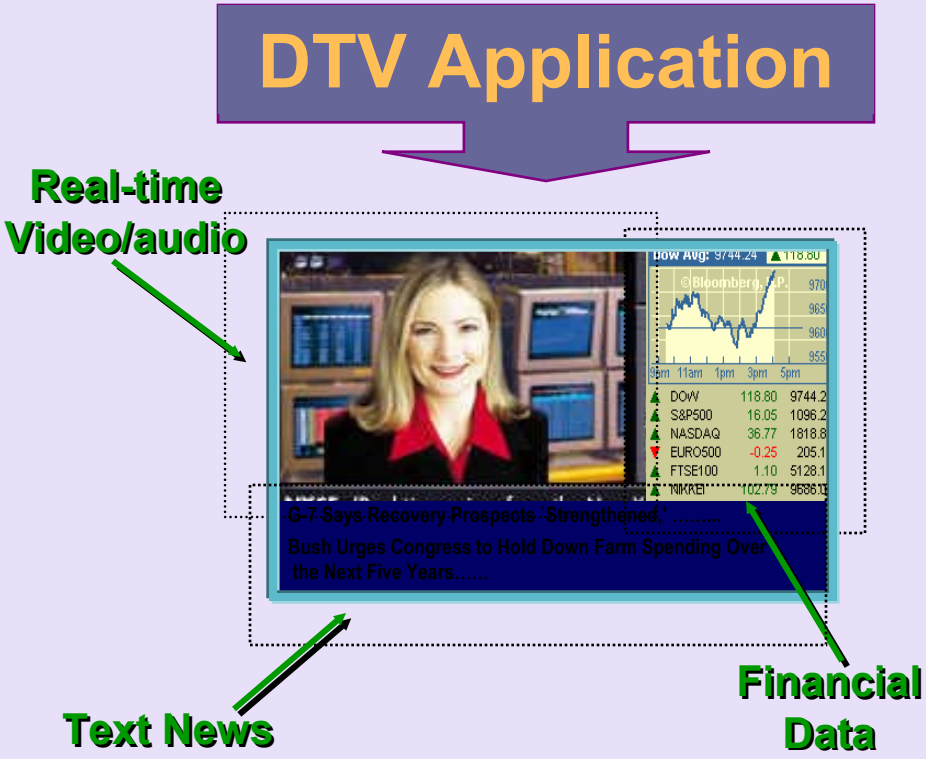
PCI Express Graphics Implementation

- Bandwidth & Scalability
- New Features & Enhancements
- Shared Memory Model
- Form factor & Power

New Features & Enhancements

- True ISOC
- Cross-segment Flexibility

Why Isochronous?



- Isochronous
  - ✓ Deadline sensitive traffic
- Digital multimedia applications
  - ✓ DTV, video on demand, Video-conferencing, Video editing
  - ✓ Peripheral connectivity (USB, 1394, e.t.c)
- Client and Comm/Server-level capability

PCI Express Isoc Capability - Enabler for Streaming Media Applications

Cross-Segment Flexibility

- Platform level
  - ✓ Reuse of x16 ports for high BW server I/O
- GPU component level
  - ✓ Multi-adapter opportunities

PCI Express Graphics Implementation

- Bandwidth & Scalability
- New Features & Enhancements
- Shared Memory Model
- Form factor & Power

Shared Memory Model

- Platform Architectural Requirements:
  - ✓ *High performance* access to system memory for *shared* non-cacheable memory buffers (command buffers and textures)
  - ✓ Robust memory management
- Implementation:
  - ✓ AGP: Platform-provided GART
  - ✓ PCI Express: IHV-provided translation

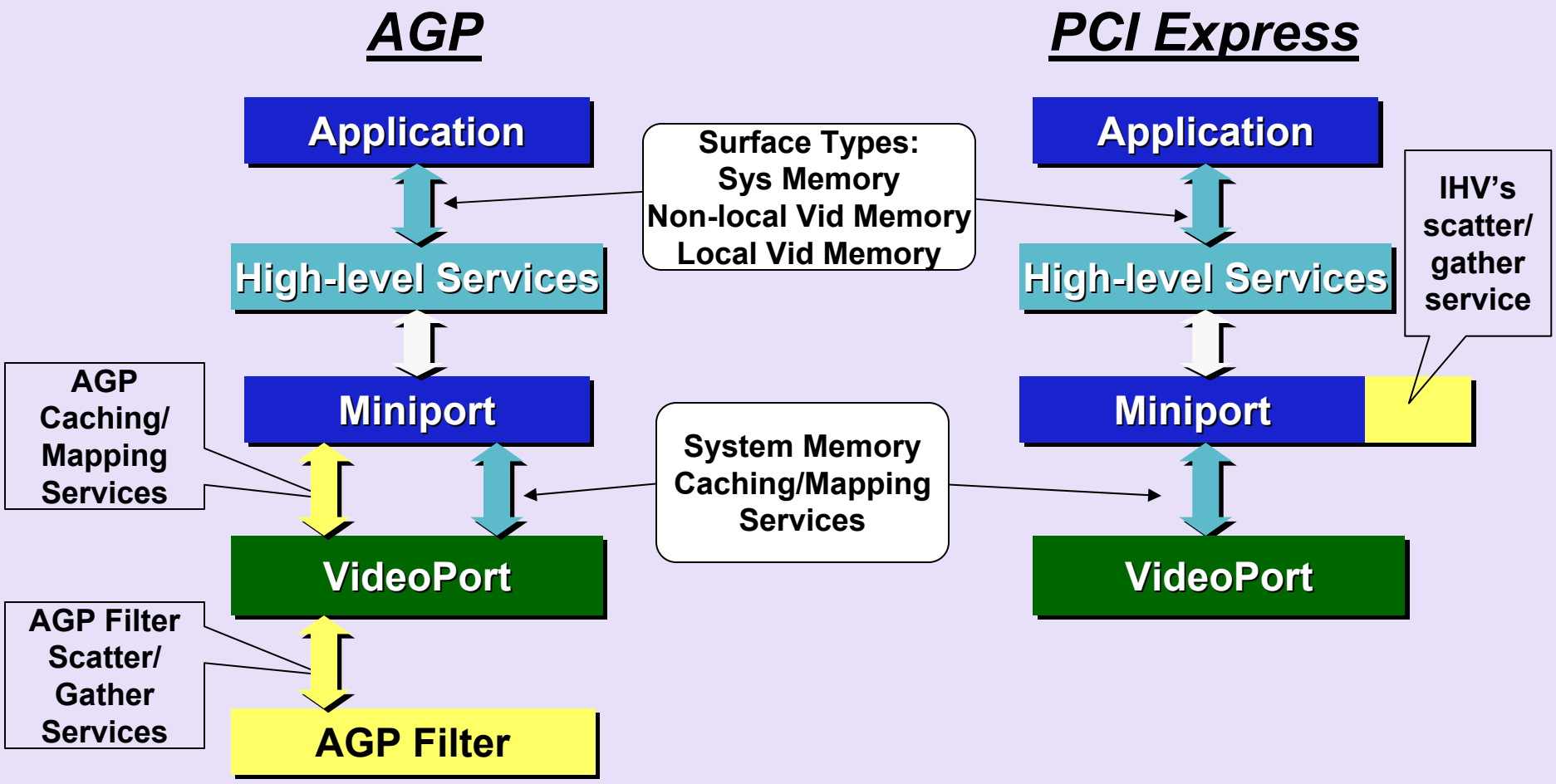
GART Transition Motivation

- Industry desire is to have a “fat pipe” to graphics
- Reduced IHV/platform dependencies
  - ✓ Improves interoperability
- Eliminate potential “graphics unique” slot requirement
- Performance optimization opportunities

SW Considerations

- Transparent to applications
  - ✓ HW abstracted from applications by today's gfx stack
  - ✓ PCI compatible software model
  - ✓ Boot Operating Systems, Configuration/Device Driver Interfaces

AGP Vs. PCI Express Graphics Stack Comparison

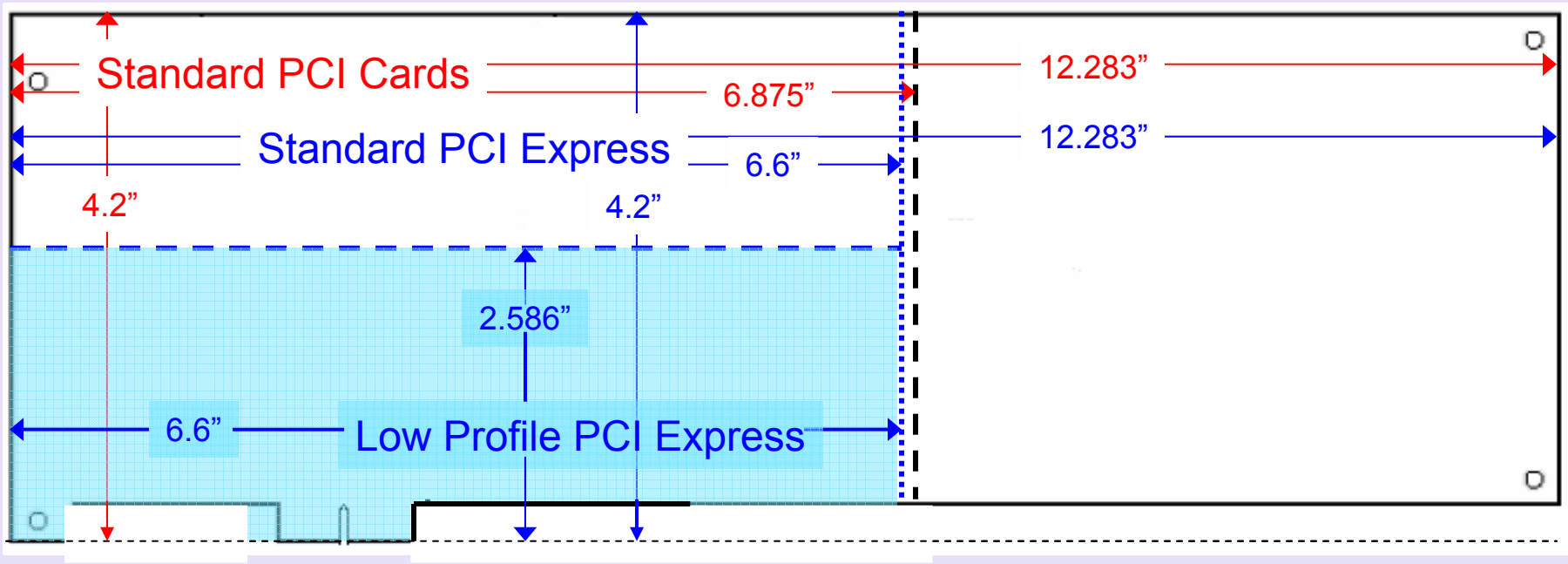


PCI Express Graphics Implementation

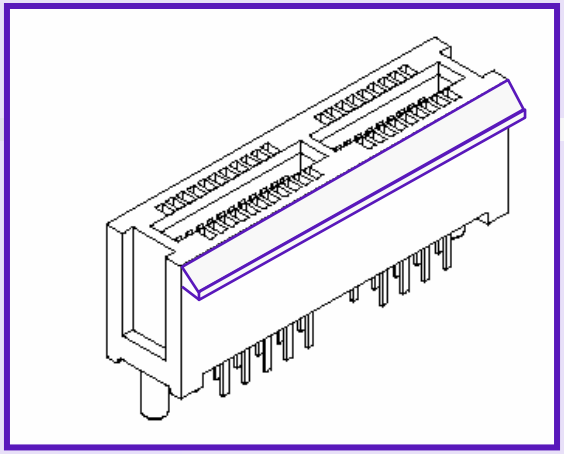
- Bandwidth & Scalability
- New Features & Enhancements
- Shared Memory Model
- Form factor & Power

PCI Express Graphics Card Form Factor

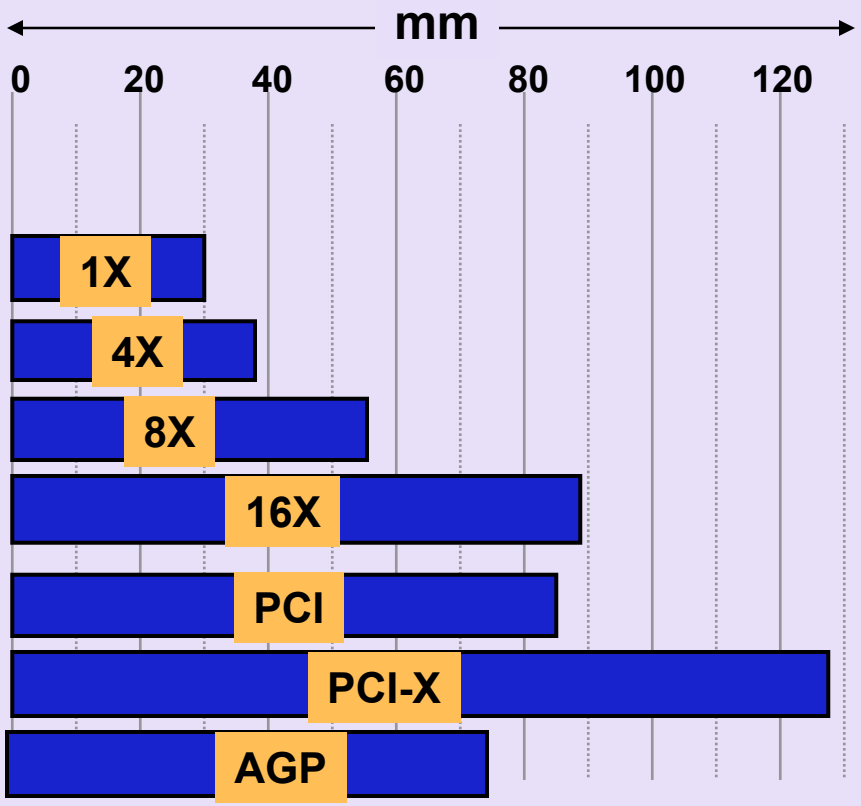
- PCI Express graphics directly leverages the standard baseline specification
  - ✓ Allowable max length of standard height card is 12.283”
    - – But not all designs will accommodate this so recommend max length is 9.5”
  - ✓ Max length of low profile is 6.6”



PCI Express Card Edge Connector



Performance Point	Connector Length: mm	(inches)
1X	25.00	(0.984)
4X	39.00	(1.535)
8X	56.00	(2.205)
16X	89.00	(3.504)
PCI	84.84	(3.400)
PCI-X	128.02	(5.040)
AGP	73.87	(2.908)



Card Interoperability

Slot / Card	x1	x4	x8	x16
x1	Yes	Yes	Yes	Yes
x4	No	Yes	Yes	Allowed
x8	No	No	Yes	Allowed
x16	No	No	No	Yes

The graphics slot will provide the most flexibility for cross segment use

- Up-plugging: Plugging a smaller link card into a larger link connector. Fully allowed and recommend to fully support.
- Down-plugging: Plugging a larger link card into a smaller link connector. Not allowed and is physically prevented.

- Down-shifting: Plugging a card into a connector that is not fully routed for all of the lanes. In general, this is not allowed. The exception is the x8 connector which the system designer may choose to route only the first four lanes. A x8 card functions as a x4 card in this scenario.

PCI Express Graphics Power and Delivery

- Enhanced power capabilities compared to AGP give PCI Express additional headroom for higher levels of graphics performance
- IDE cable power is **NOT** supported
- Multiple distinct levels
  - ✓ 25W for low profile cards
    - Power from slot only
  - ✓ 75W for standard height cards
    - Power from slot only
  - ✓ 150W for standard height high-end cards
    - Power from slot and NEW PCI Express graphics power connector

PCI Express Graphics CEM Power

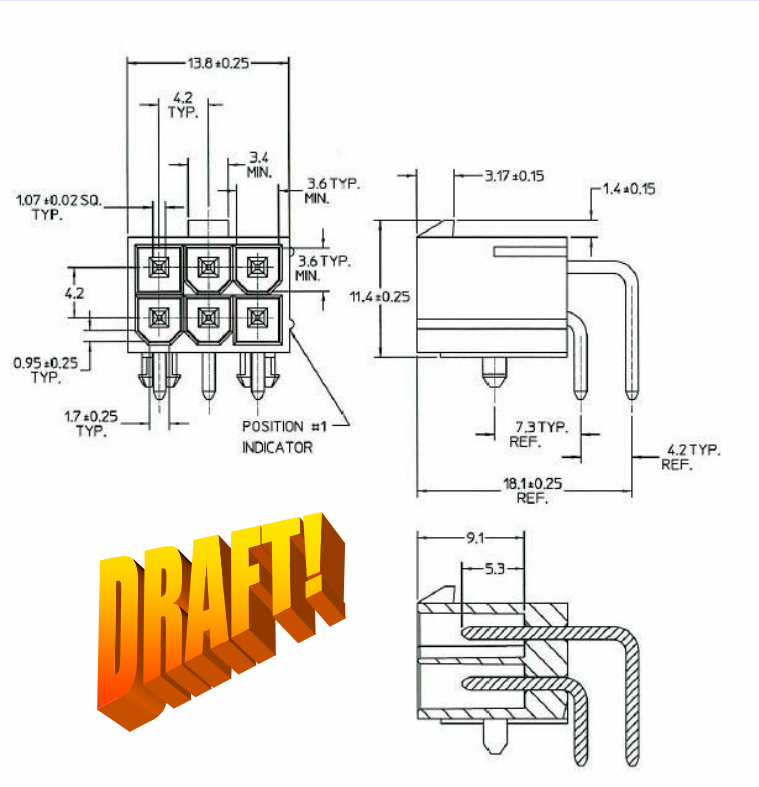
# Links / Power Rail	25W slot (low profile cards)	75W slot
3.3V ±0.3V	3A max	3A max
12V ±1.0V	2.1A max	5.5A max
3.3Vaux ±0.3V	375mA max	375mA max

PCI Express High End Graphics Specification

- Objective of PCI Express High-End Graphics:
  - ✓ Continue to enable the highest end graphics on a PC
  - ✓ Standardize power level and distribution
- Provides an additional 75W of power to graphics
  - ✓ Via a dedicated and unique power supply connector
    - 2x3 connector that must be directly cabled to the PSU
    - Does **not** support usage of any dongles or adapters
  - ✓ Graphics now has access to an aggregate of 150W
- Specification is quickly approaching v1.0

Supplemental power connector for High End Graphics

- PRELIMINARY drawing of the right-angle, through-hole PCB connector shown at right
  - ✓ 3 +12V pins, 2 Ground pins, 1 sense pin (tied to Ground in PSU or cable connector)
    - – The card MUST keep this 12V rail separate from the 12V rail from the slot!
  - ✓ 8A/pin max current
  - ✓ Polarized
  - ✓ Retention lock
- Cable will use 18AWG wire



Dimensions subject to change - do NOT proceed with designs!

12V Power key points for HE Gfx cards

- The system power supply must provide separate 12V rails between the card edge power and the supplemental cable power
- The graphics card must maintain separate 12V planes
- The graphics card must be able to handle the extreme range of 12V tolerance between rails
  - ✓ i.e. one rail could be at +9% tolerance while the other rail could be at -9%

HE Gfx card retention

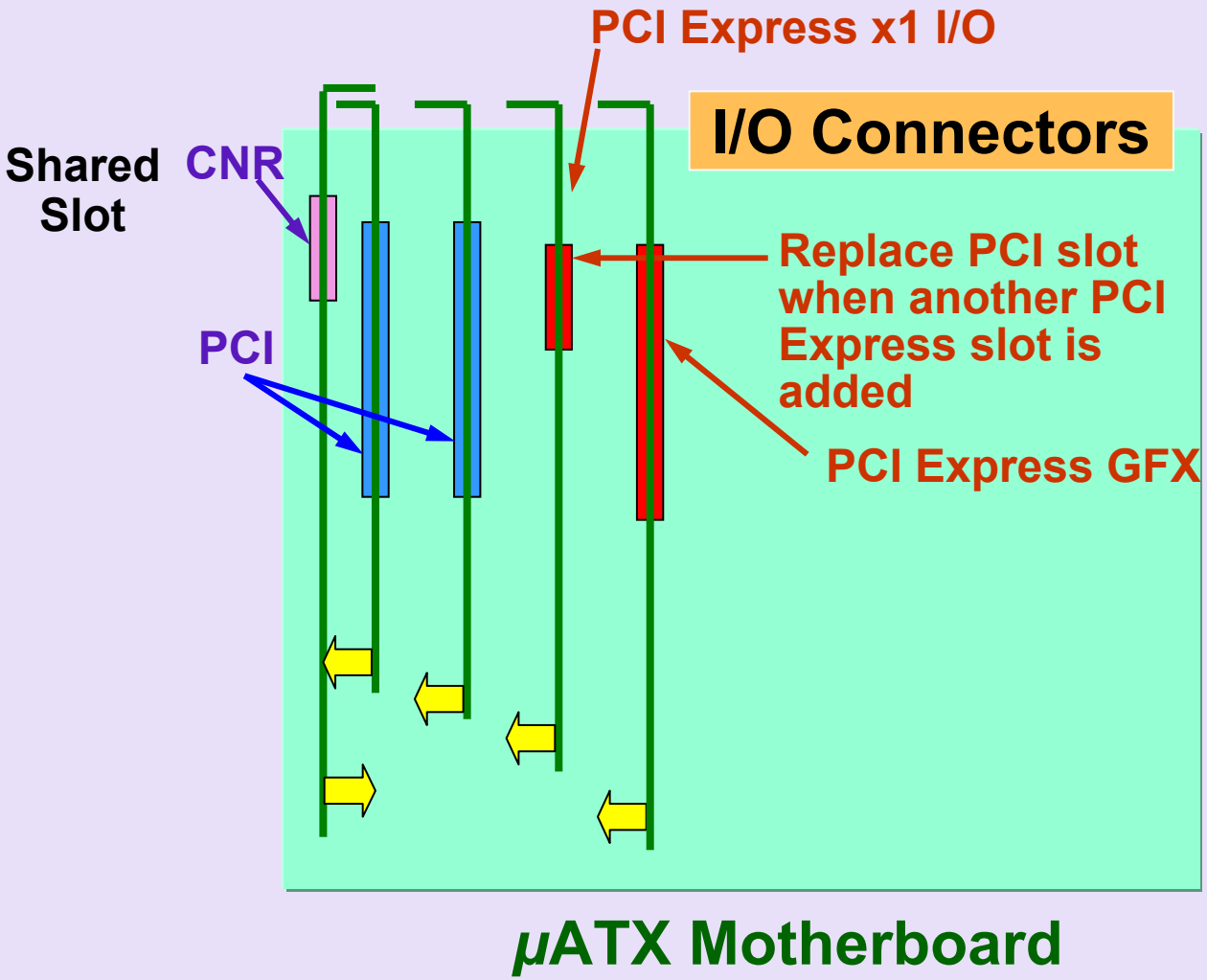
- Additional card retention and support is required for cards that are greater than 350 grams in mass!
- The “hockey stick” feature defined in the CEM spec is optional for a HE Gfx card.
  - ✓ This mechanism is insufficient for high mass cards
  - ✓ It interferes with certain add-in card thermal solutions

Agenda – PCI Express Graphics

- Introduction to Graphics Interconnects
- PCI Express Graphics Implementation
- Transition strategy
- Summary & Call to Action

PCI Express Evolutionary Strategy

- "Evolutionary" aka PCI Express, can directly use existing ATX chassis
- PCI Express can be implemented in ATX/uATX designs
- Increased routing and component area available with smaller PCI Express x1 connector



Agenda – PCI Express Graphics

- Introduction to Graphics Interconnects
- PCI Express Graphics Implementation
- Transition strategy
- Call to Action & Summary

Call to Action

- Ready product roadmaps to intercept launch
- Ensure designs meet PCI-SIG compliance
  - ✓ Especially critical for power distribution
- Utilize the PCI-SIG for specifications and support

Question & Answers

Thank you for attending the
2004 PCI-SIG Developers Conference.

For more information please go to
www.pcisig.com

Backup material

Interface Comparison Summary

	AGP8x	*PCI Express (x16)*	*Notes*
Peak BW	2GB/sec (3rd gen)	4 GB/sec (1st gen)	
Peak Concurrent BW	2GB/sec	8 GB/sec	
BW/Pin	~18MB/sec/pin	~100MB/sec/pin	
Request Size Range	8B-64B	4B-4096B	
Request pipe depth max	32	256	PCI Express extensible to 2K using PFN
#independent streams	up to 3 (PCI, LP, ISOC)	up to 8 (VC0-7)	
ISOC	Partial	Full	AGP lacks arb / regulation mechanism Restrictive topologies

Interface Comparison Summary

	AGP8x	*PCI Express (x16)*	*Notes*
Multi-head	N*	Y	*AGP spec did not preclude, but implementation complexities hindered adoption
Shared System memory	Y (GART)	Y (OS/Driver)	Application transparent
Flow Control	Request: source controlled credit-based Data: destination controlled (RBF#/WBF#)	Req/data: source controlled credit-based	
Ordering	Stream-dependent Relaxed rules LP/ISOC	Stream-dependent PCI-X relaxed rules	
Physical I/F	Single-ended, Source-synchronous clk	Differential, embedded clk	
Power	25W (50/110W for AGP-pro)	75W (150W for PCIe High End Graphics)	40W under review in WG


PCI
SIG™