IBM System p

POWER Roadmap

Bradley McCredie
IBM Systems &Technology Group, Development
IBM Fellow

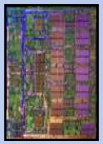
IBM Systems

IBM POWER SYSTEMS

Consistent Predictable Delivery

2001 — POWER4

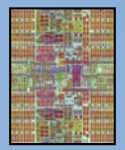
2003 — POWER4+



2004 — POWER5

2006 — POWER5+

2007 — POWER6



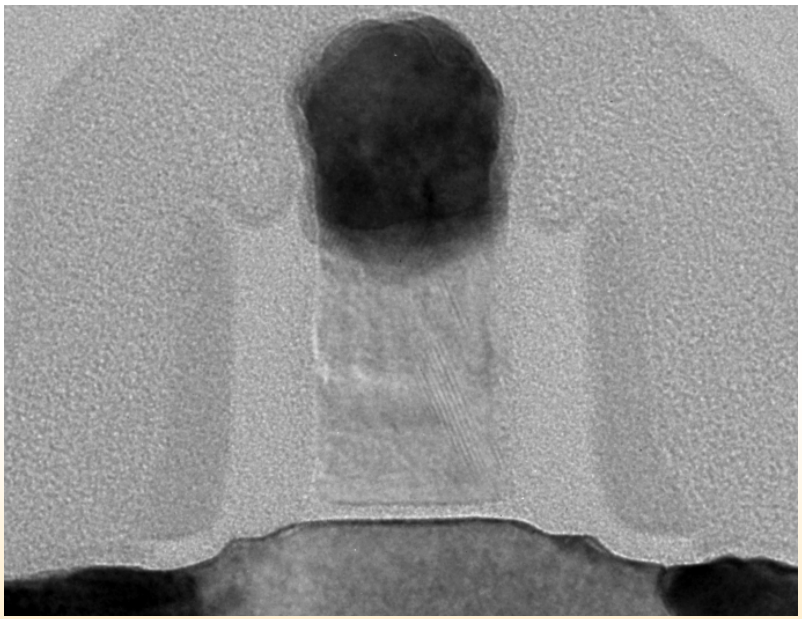
POWER6™

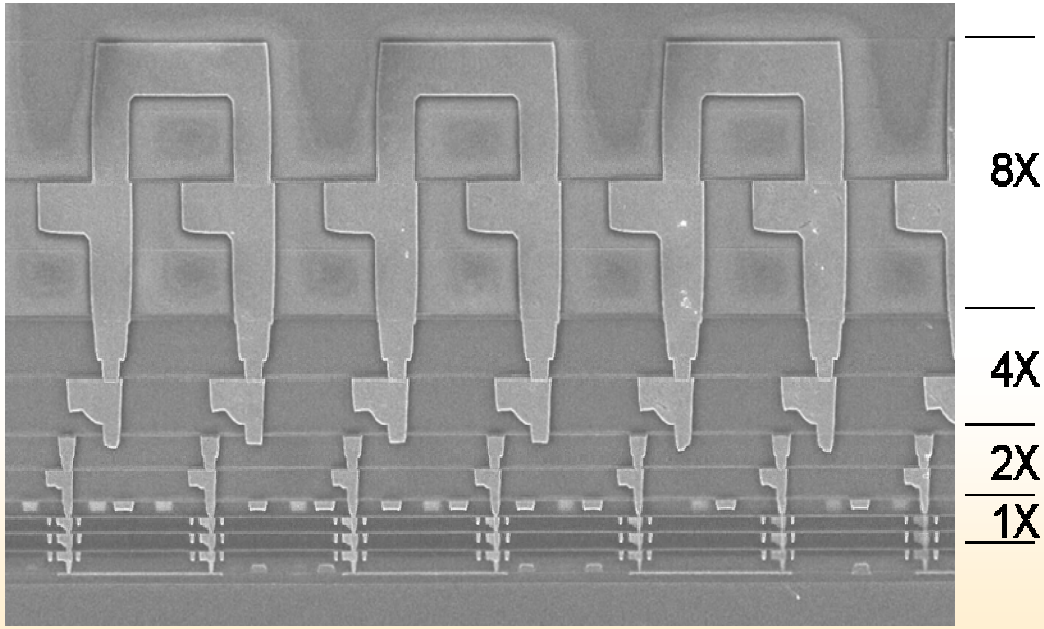
Innovations in all areas of system design:

- **Technology**
- **Performance**
- **Flexibility**
- **RAS**
- **Features and function**
- **System Energy Management**
- **Current development Status**

Key 65nm technology features

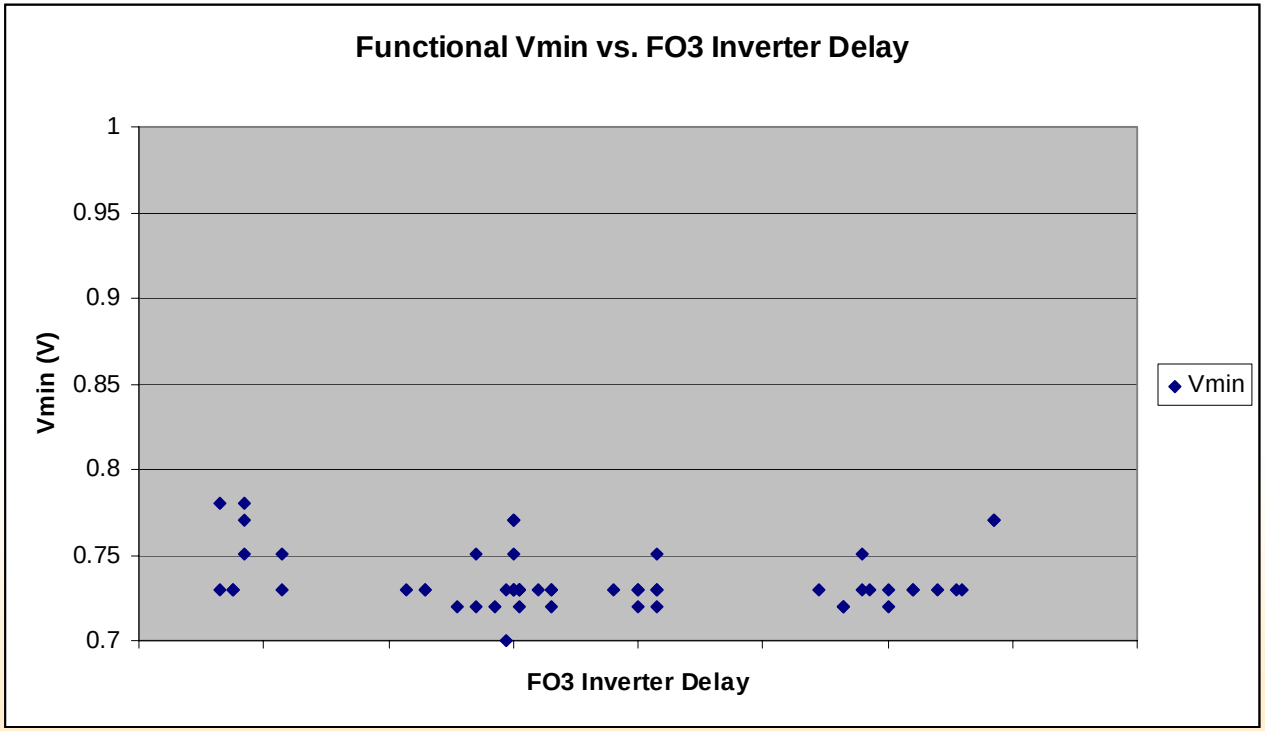
- **30% performance improvement over 90nm at constant power**
  - **65nm SOI technology with 10 levels of metal (low-k dielectric on first 8 levels)**
  - **Pitches: 180nm M1, 200nm M2 to M4, 400nm M5-M6, 800nm M7-M8, and 1600nm M9-M10**
  - **250nm contacted gate pitch**
  - **35nm gate length, 1.05nm gate dielectric thickness**
  - **dual stress liners**
  - **0.65um2 high-performance SRAM cell**





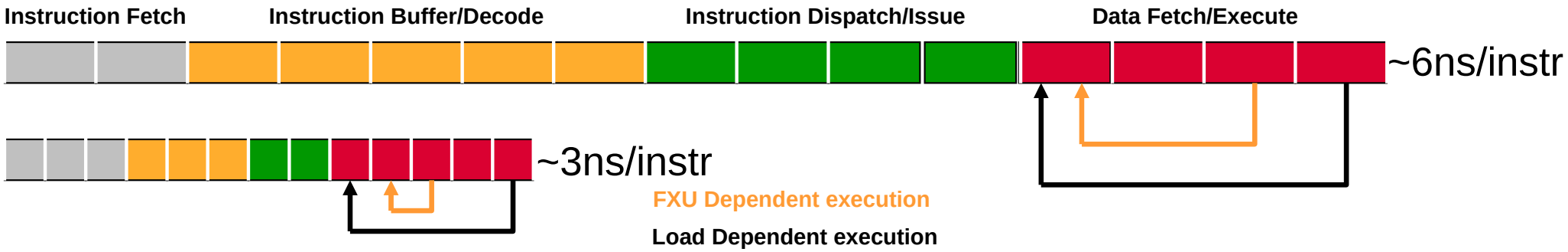
Technology power/performance features

- **Additional technology features to optimize low power design:**
  - Tailored transistors for ultra low power
  - Tailored array cells for high performance and high density applications
  - All array cells use separate voltage supply to enable low voltage application
  - Broad voltage range of technology operation for low power and high performance application



POWER6 Core

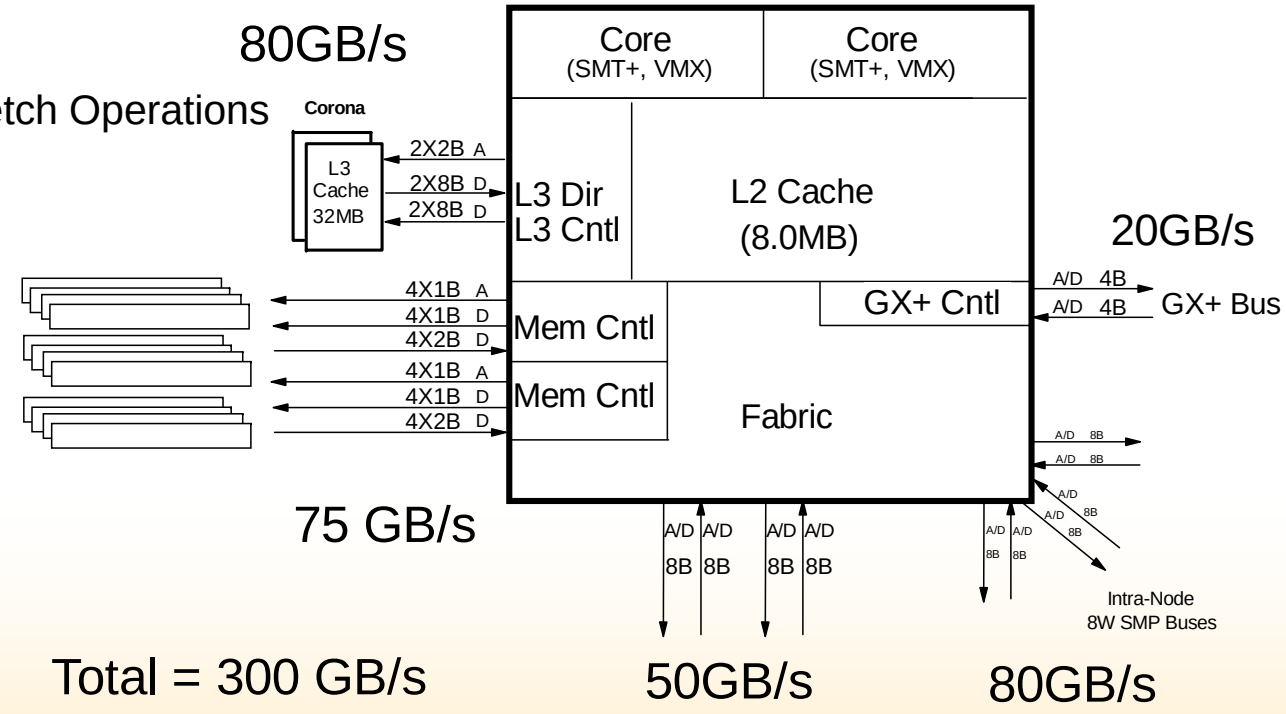
- **POWER6 processor is ~2X frequency of POWER5 (4-5GHz)**
- **POWER6 instruction pipeline depth equivalent to POWER5**
  - Minimize power
  - Scale performance with frequency



- **POWER6 Extends functionality of POWER5 Core**
  - 64K I Cache, 64K D Cache, 2 FXU, 2 FPU, 1 Branch execution unit
  - Two way SMT with 7 instruction dispatch from 2 threads (maximum of 5 instructions per thread)
  - Decimal Unit
  - VMX Unit
  - Recovery Unit

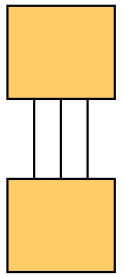
POWER6 scales chip capabilities with core performance

- **Cache highlights**
  - 4MB Private L2 Cache per Core
  - 32MB Non-sectored L3 Cache per chip
- **Fabric highlights**
  - Three Intra-Node SMP buses for 8-way Node
  - Two Inter-Node SMP buses for up to 8 Nodes
  - Multiplexed Address/Data SMP buses
- **New prefetching capabilities**
  - Coherent Multi-Cacheline Data Prefetch Operations
  - Prefetching on stores

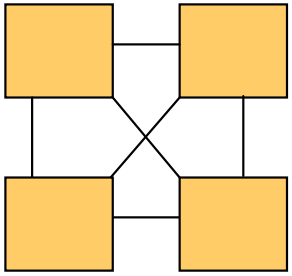


Flex System to optimize low end to high end server designs

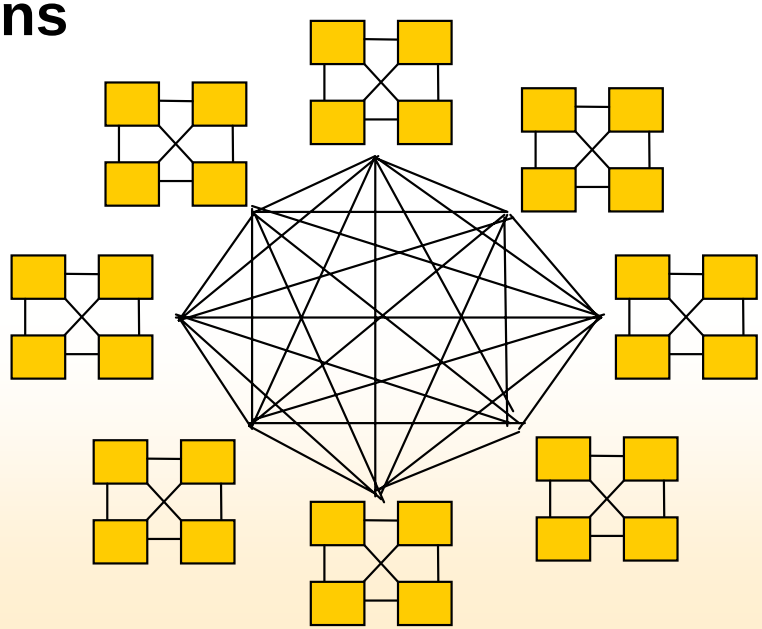
- **SMP busses can be configured in two modes**
  - Cost/performance trade-offs
  - On node busses are 8B or 2B
  - Off node busses are 8B or 4B
- **Numerous memory controller BW options**
  - 1 or 2 memory controllers are available
  - Memory controllers can be configured to full width or ½ width
- **L3 cache is supported in three configurations**
  - On module High Bandwidth configuration
  - Optional off module configuration
  - No L3 option
- **Fully interconnected two-tier SMP fabric**
  - Reduced latencies vs. POWER5
  - New two tier memory coherency protocol



2 socket



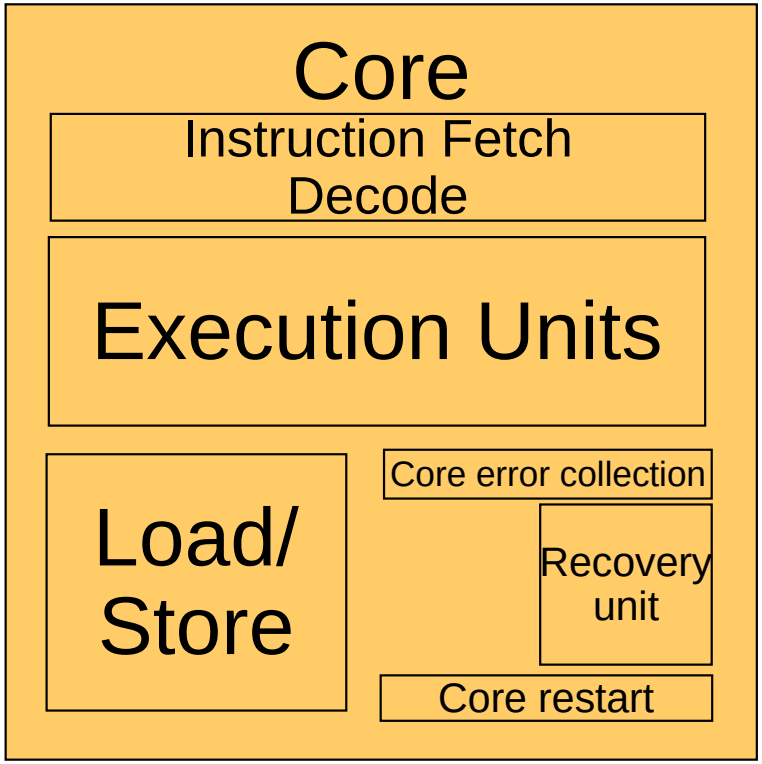
4 socket



32 socket

Bullet-proof computing

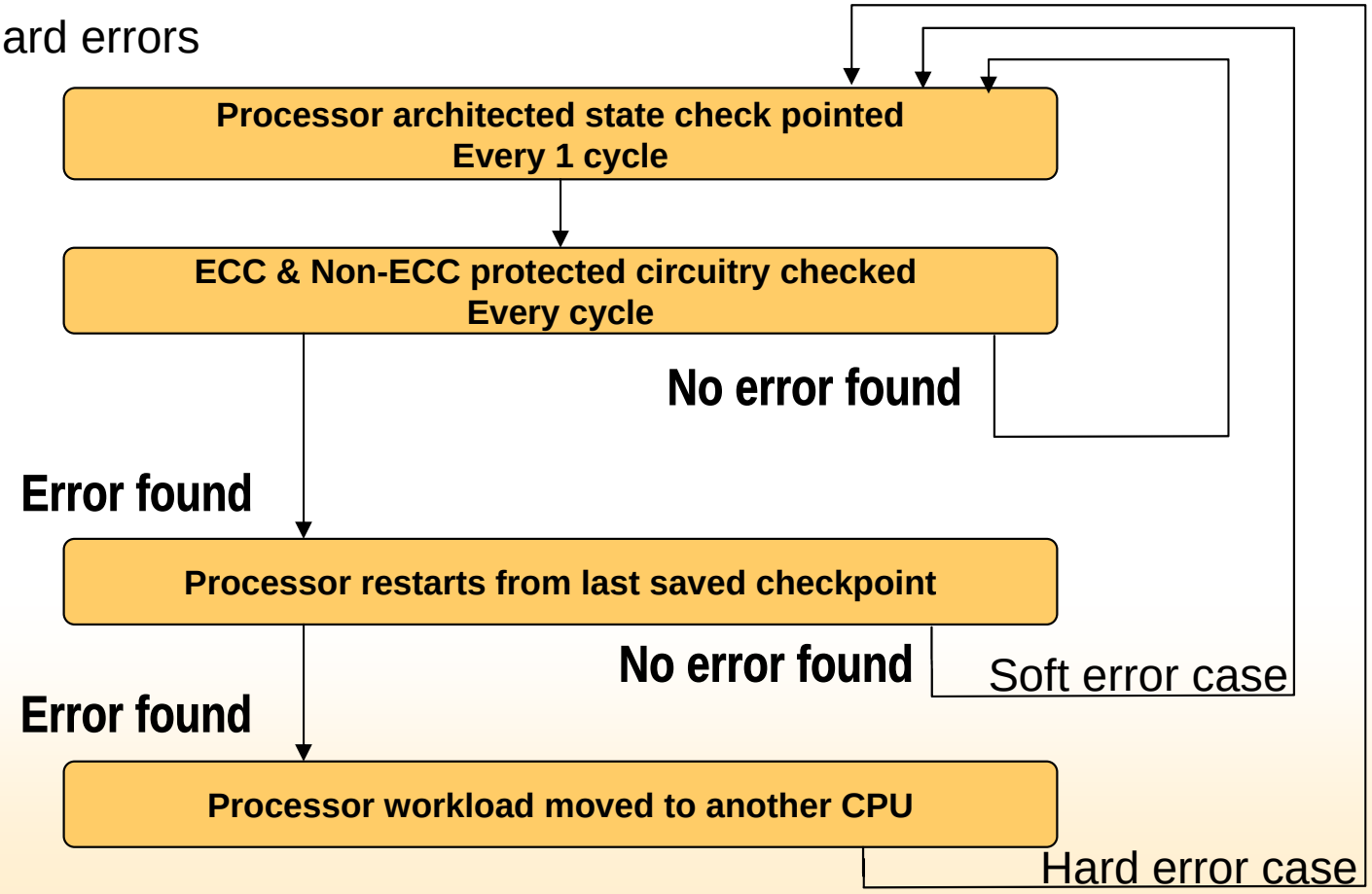
- **Recovery Capability**
  - Array error
    - Error correction (ECC)
    - Arrays with parity
      - Processor restarts
  - Instruction flow and Data flow Error
    - Processor restarts
  - Control Error
    - Processor restarts



- **System Resiliency**
  - Processor states are check pointed and protected with ECC
  - Processor states can be moved from one processor to another upon unsuccessful recovery restart

Bullet-proof computing

- **System RAS with recovery unit**
  - Every measure possible taken to preserve application execution
  - Retry soft errors
  - Change hardware for hard errors



Features and functions

IN-CORE HARDWARE ACCELERATORS
Decimal Floating point and Altivec (VMX)

VIRTUALIZATION ENHANCEMENTS

3rd GENERATION MULTI-THREADING

Decimal Floating Point accelerators

- **Binary floating-point unsuitable for commercial or human-centric applications**
  - Cannot meet legal and financial requirements
- **Survey of numeric data in commercial databases is largely decimal data**
  - 55% of numeric data in databases is BCD data
  - The next 43% of data is integers, often held as decimal integers
- **Example: performance improvement of decimal hardware vs. decimal software**
  - Telco billing application -- 1 million calls (2 minutes) read from file, priced, taxed, and printed:

	Java BigDecimal w/ DEC number	C, C# packages	Integer hand-tuned
% execution time in decimal operations	93.2%	72 – 78%	45%
Speedup with hardware DFP*	< 7X	4X	2X

*** IBM projection**

Decimal Floating-Point Unit

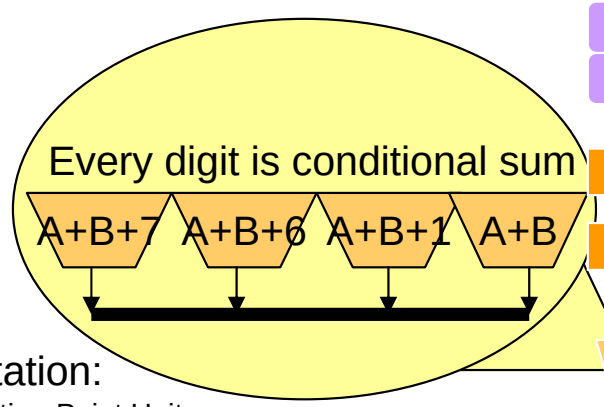
Architecture:

Added ~50 instruction to POWER ISA:

- Add, Multiply, Divide
- Conversions (to/from Integer, BCD & DFP)
- Insert Exponent (scaling)
- Quantize (single unit)
- Reround (round to less precision once)

Supports IEEE 754R standard which is going to ballot this month

Adder
2 by 18 digit Mult
1 by 36 digit Quad
2 cycle Pipelineable



Minimal Hardware Implementation:

Separate unit from Binary Floating-Point Unit
Reuses BFUs Register File (FPRs) to minimize area
Reuses FPSCR (status and control)

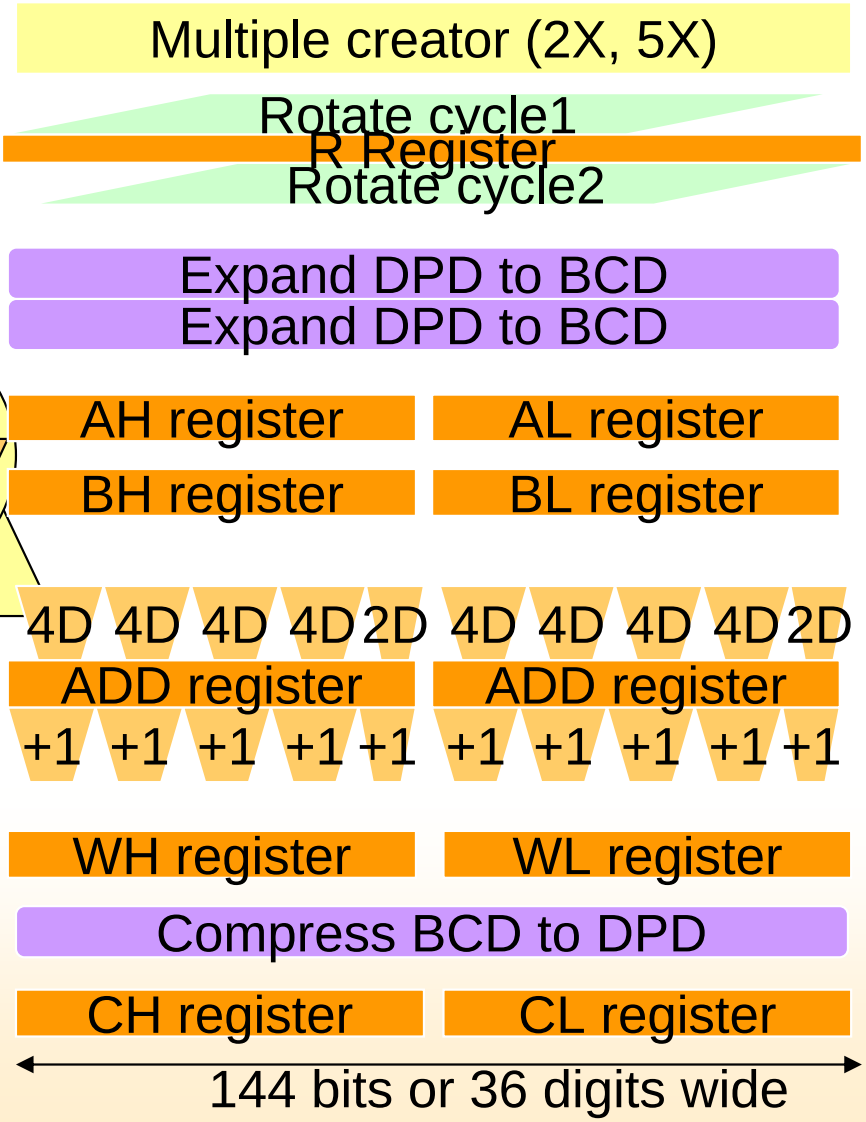
- Separate Decimal Rounding Mode (Bankers rounding)

Quadprecision dataflow

- 34 digit + 2 guard, 144 bit wide, compresses to 128bit in memory

Both 34 digit and 16 digit datatypes supported in hardware
36 digit adder

- 2 cycle latency
- 1 cycle throughput at CPU cycle



Standards & software rallying around DFP standardization

- **Numerous software and standards activities in flight inside and outside IBM**
  - Java BigDecimal (compatible with 754r)
  - C# and .Net ECMA and ISO standards arithmetic changed to match 745r decimal128
  - XML Schema 1.1 draft now has *pDecimal compatible with 754r*
  - ISO C and C++ are jointly adding decimal floating-point as first-class primitive types
  - GCC almost complete
  - Cobal is adding a datatype to support 754r
  - ANSI/ISO SQL … new types accepted in principle (draft about to be submitted)
  - Strong support expressed by Microsoft, SHARE, academia, SAP and many others

Other core features

- **Virtualization enhancements**
  - Hardware support for increased virtualization granularity
    - Up to 1024 partitions are supported
  - Support virtual partition of memory
    - Memory can be reconfigured and moved for partitions (memory packing)
    - Allow concurrent maintenance by reconfigure and isolate memory failures
  - Support virtual page key protection
    - Data page protected by keys to prevent unauthorized access
- **Simultaneous Multithreading Enhancements**
  - SMT provides excellent power performance
    - Reuse of existing transistors vs. performance from additional transistors
  - Second thread/core delivers tremendous throughput
    - 55% performance on OLTP
    - 40% performance on integer application
  - Performance improvements achieved with architectural enhancements
    - Enhanced cache sizes and associativity to minimize thrashing effect from the 2 threads
    - Improve dispatch bandwidth for SMT

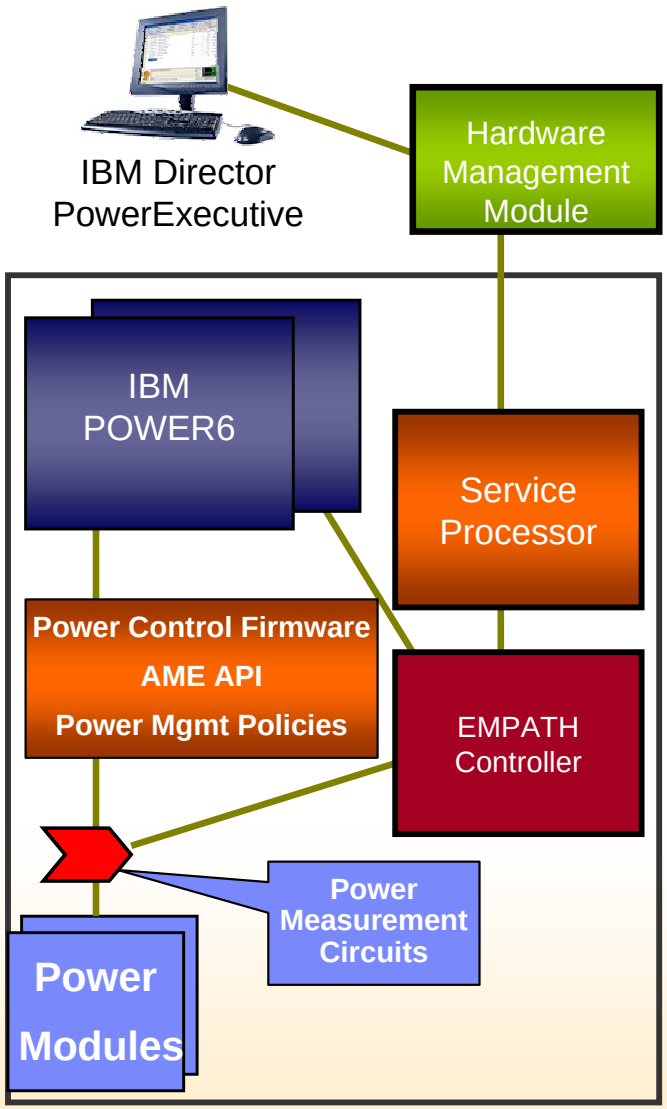
PowerExecutive™ Extensions for POWER6 Energy Management Policies

Example Energy Management Policies:

- **Energy cost management**
  - Monitor System workloads/power consumption
    - If: System utilization reduces reduce system power/performance
    - If: Multiple Systems go below utilization threshold consolidate workloads
    - If: System power budgets exceed allocation cap power
- **Acoustic optimization**
  - Monitor Systems temperature
    - If system temperatures go below threshold reduce fan speeds
- **Performance optimization**
  - Monitor system temperature/power consumption
    - If temperatures/power consumption go below threshold increase performance

Energy Management Policies Enable Customers To Maximize The Compute Capability Of Their Datacenter Or Minimize Energy Costs

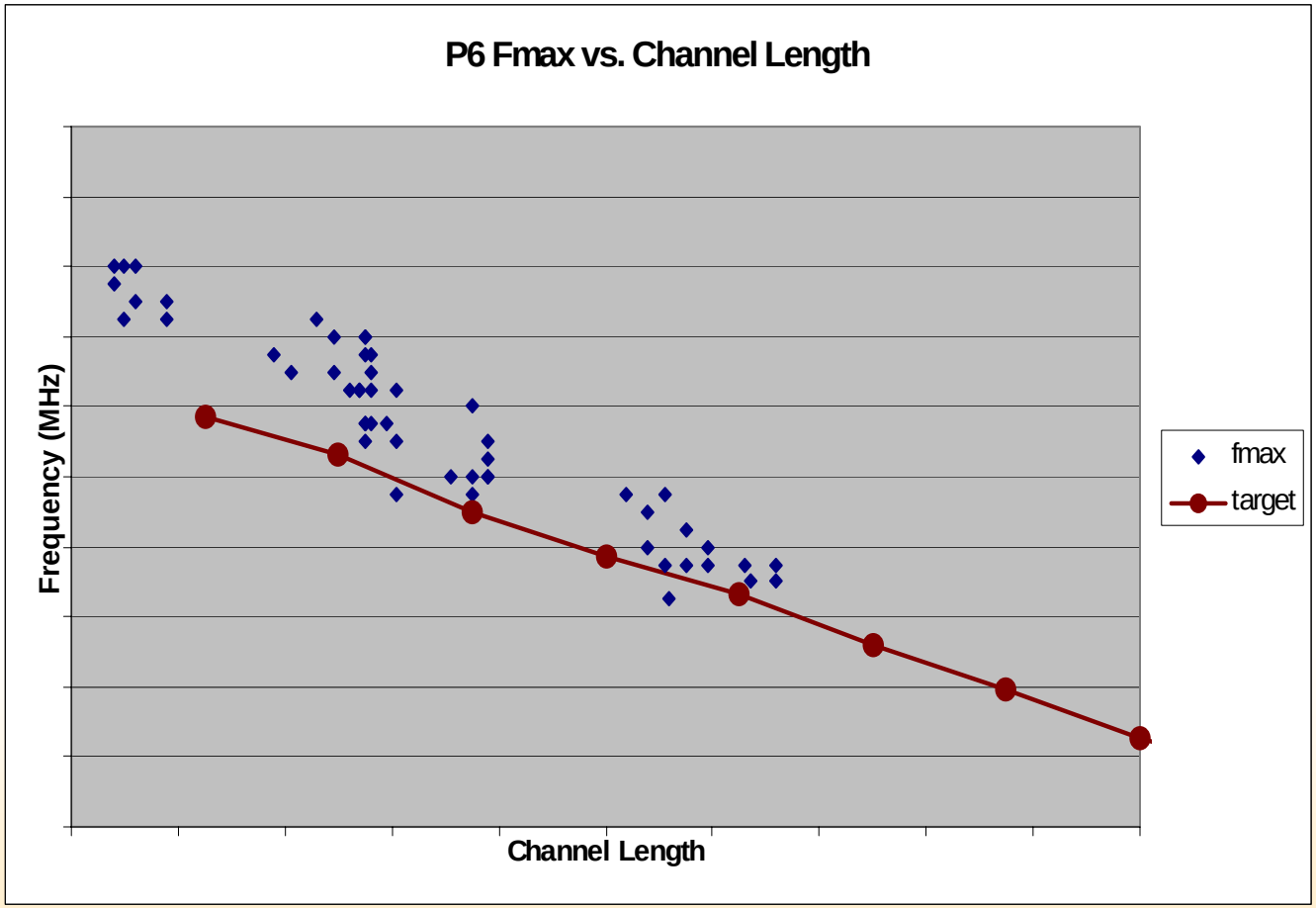
POWER6 EMPATH System Control
Extended System Functions For PowerExecutive Policies



- **Thermal / Power Measurement**
  - Read thermal data from processor chip thermal sensors
  - Measure power data from system level sensors
  - Report data via PowerExecutive
- **Power Capping**
  - Use of Hardware controls to keep system power under a specified limit
- **Power Saving**
  - Operation at reduced power when workload and policy allows
    - Can be a static policy (e.g. overnight reduction)
    - Can be dynamic (when absolute max performance is not always required)
- **System health monitoring**
  - Use of hardware sensors to ensure system is operating within safe predefined bounds
- **Performance-Aware Power Management**
  - Use of dedicated performance counters to guide power and thermal management tradeoffs

POWER6 implementation status

- **Production-level chip design is in the lab**
- **Numerous systems are in various stages of bringup, debug and test**
  - High End
  - Midrange
  - Low End
  - Cluster/Scientific
- **On target for mid ’07 GA**



Summary & Conclusions

- **POWER6 development and delivery is on schedule**
- **POWER6 doubles frequency and bandwidth of POWER5™**
  - Same pipe depth
  - Same power envelope
- **POWER6 scales chip/system performance with core performance**
- **POWER6 provides new capabilities**
  - Decimal Floating Point
  - Processor recovery
- **System p™ will begin delivery of system power management with POWER6**
- **POWER6 is on track to deliver high frequency capabilities on schedule**