Smarter Systems for a
Smarter Planet

IBM zNext –

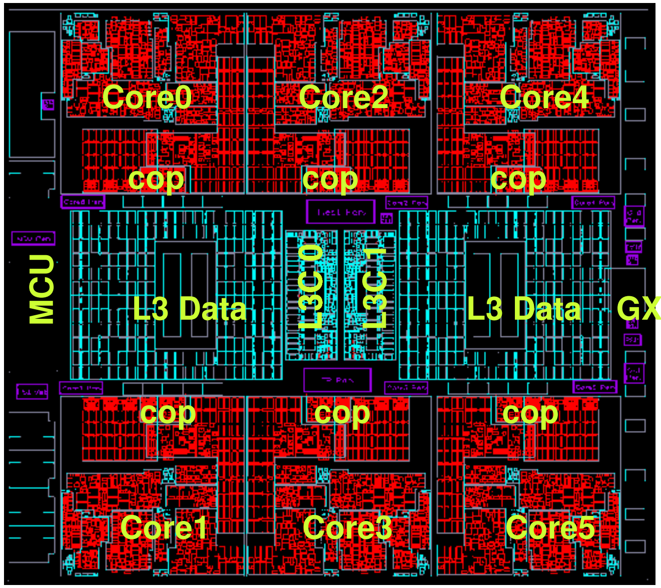
The 3rd Generation High Frequency Microprocessor Chip

Chung-Lung (Kevin) Shum

Senior Technical Staff Member, System z Processor Development, Systems & Technology Group, IBM Corp.

© 2012 IBM Corporation

zNext PU Chip Overview



- IBM mainframe microprocessor chip for the next generation of System z servers

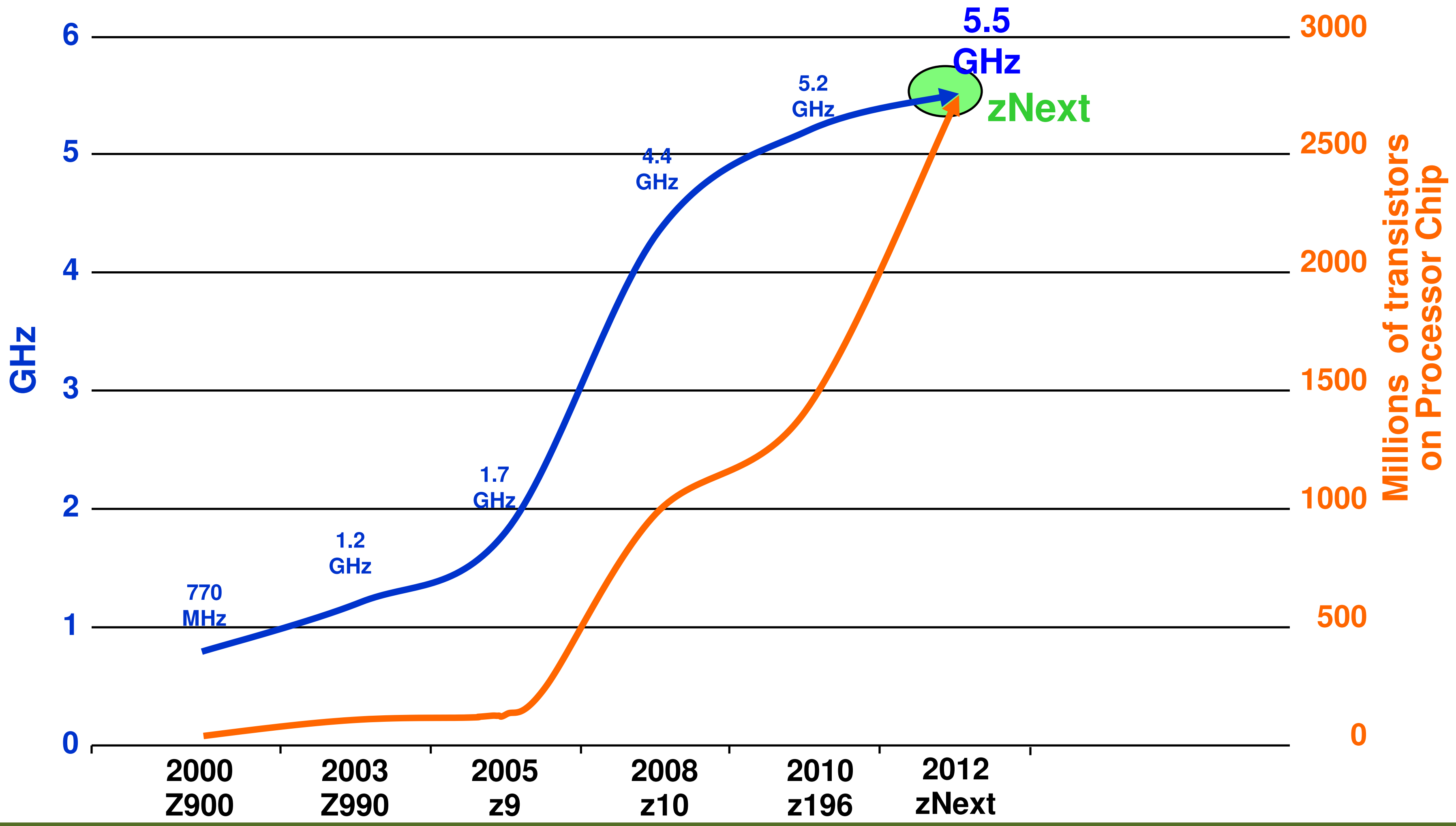
- 32nm SOI technology
  - 597 mm2 (23.7mm x 25.2mm)
  - 15 layers of metal, 7.68 miles of wire
  - 2.75 Billion transistors
  - I/Os: 10000+ Power, 1071 Signal
    - SMP connections to external Hub chip (SC)
    - I/O Bus Controller (GX)
    - Memory Controller (MC) with **prefetching**

- Chip Features (vs. z196)
  - **6 new cores** per chip (vs. 4)
  - Core-**Dedicated** (vs. shared) Co-Processors
  - **48 MB EDRAM** on-chip shared L3 (2x)

- Processor Core Features
  - 2nd Generation out-of-order design
  - Speed & feed improvements
  - Microarchitecture innovations
  - Architecture extensions for software exploitations, e.g.,
    - Hardware Transactional Memory
    - Runtime instrumentation

© 2012 IBM Corporation

Speed: Higher Operating Frequency



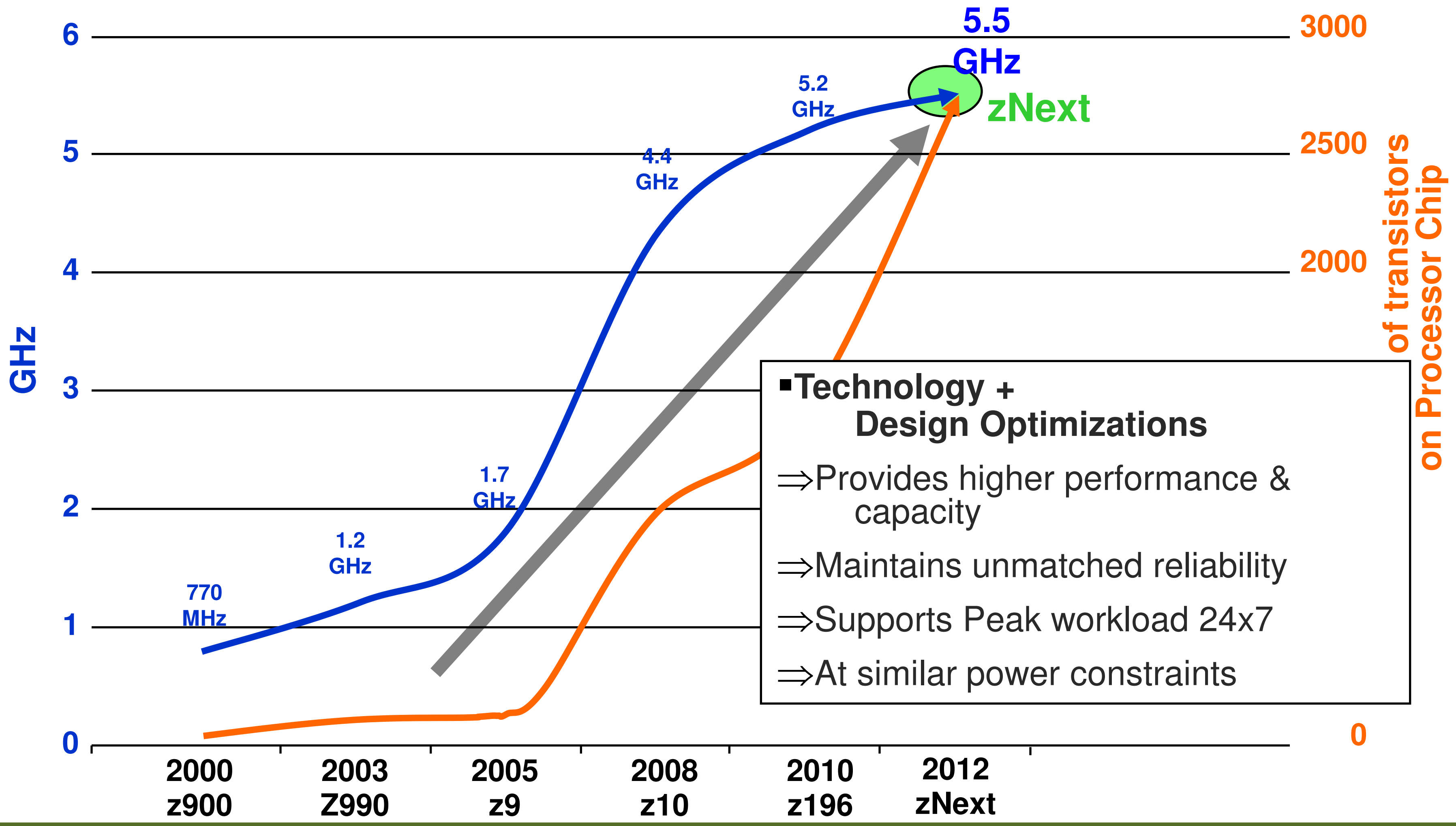
- **z900** – Full 64-bit z/Architecture
- **z990** – Superscalar CISC pipeline
- **z9** – System level scaling

- **z10** – Deep Pipeline, Arch. extensions
- **z196** – Out-Of-Order (OOO), Additional Architectural Extensions

- **zNext** – OOO+, Architectural Extensions, Enablement for new Software Paradigms

© 2012 IBM Corporation

Speed: Higher Operating Frequency



- **Technology + Design Optimizations**

⇒Provides higher performance & capacity

⇒Maintains unmatched reliability

⇒Supports Peak workload 24x7

⇒At similar power constraints

- **z900** – Full 64-bit z/Architecture
- **z990** – Superscalar CISC pipeline
- **z9** – System level scaling

- **z10** – Deep Pipeline, Arch. extensions
- **z196** – Out-Of-Order (OOO), Additional Architectural Extensions

- **zNext** – OOO+, Architectural Extensions, Enablement for new Software Paradigms

© 2012 IBM Corporation

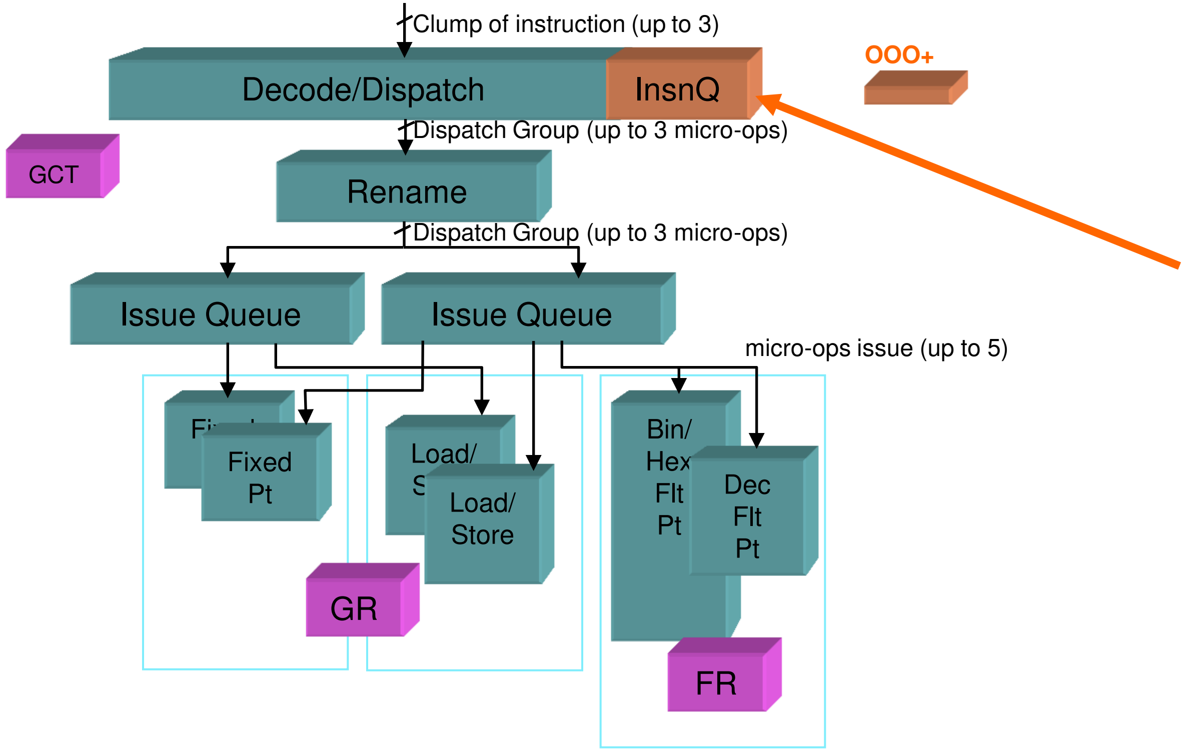
Feed Improvements: Maximizing Out-of-Order Window

- Improved dispatch grouping efficiencies

  ➔ More instructions per group

  - Reduced cracked instructions overhead
  - Increased branches per group to 2
  - Added Instruction Queue (InsnQ) for re-clumping

*clumps – parcel of instructions delivered from instruction fetching


Clump of instruction (up to 3)
OOO+
Decode/Dispatch
InsnQ
GCT
Dispatch Group (up to 3 micro-ops)
Rename
Dispatch Group (up to 3 micro-ops)
Issue Queue
Issue Queue
micro-ops issue (up to 5)
Fixed Pt
Load/ Store
Bin/ Hex Flt Pt
Dec Flt Pt
GR
FR

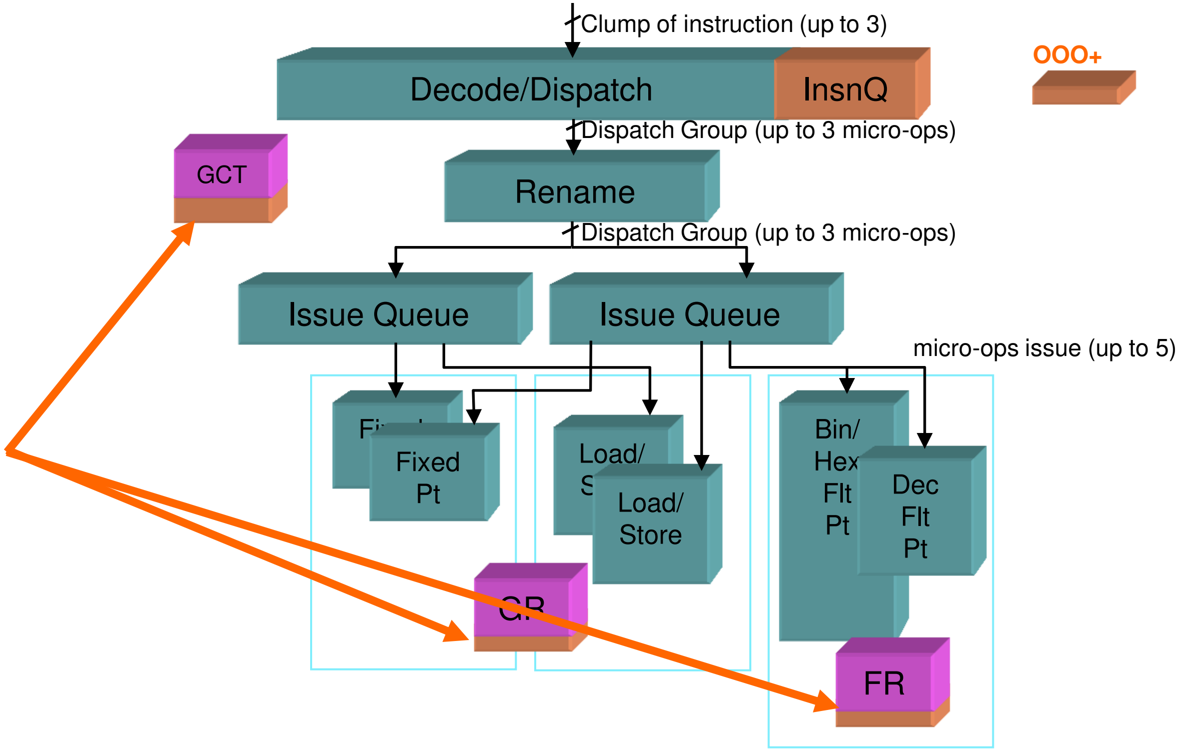
© 2012 IBM Corporation

Feed Improvements: Maximizing Out-of-Order Window

- Increased out-of-order resources

  ➔ More out-of-order groups

  – Multi-grouped instructions speculative completion
  – Increased Global Completion Table (GCT) entries to 30x3 (+25%)
  – Increased usable physical GR entries to 80 (+25%)
  – Increased physical FR entries to 64 (+33%)


Clump of instruction (up to 3)
OOO+
Decode/Dispatch
InsnQ
Dispatch Group (up to 3 micro-ops)
GCT
Rename
Dispatch Group (up to 3 micro-ops)
Issue Queue
Issue Queue
micro-ops issue (up to 5)
Fixed Pt
Load/Store
Bin/Hex Flt Pt
Dec Flt Pt
GR
FR

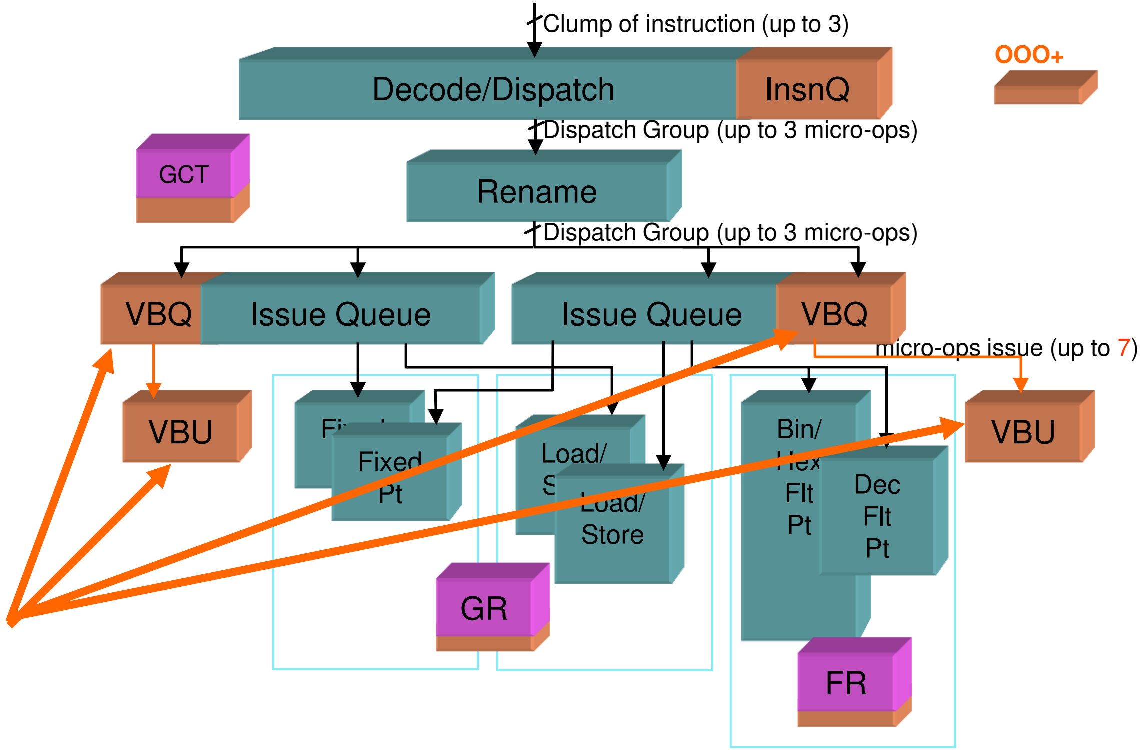
© 2012 IBM Corporation

Feed Improvements: Maximizing Out-of-Order Window

- Increased execution bandwidth

  ➔ More instructions issued per cycle

  – Added Virtual Branch Queue (VBQ) for relative branch queuing
  – Added Virtual Branch Unit (VBU) for relative branch execution
  – Increased effective issue queue size to 32x2 (+60%)
  – Increased issue bandwidth per cycle to 7 (+40%)


Clump of instruction (up to 3)
OOO+
Decode/Dispatch
InsnQ
Dispatch Group (up to 3 micro-ops)
GCT
Rename
Dispatch Group (up to 3 micro-ops)
VBQ
Issue Queue
Issue Queue
VBQ
micro-ops issue (up to 7)
VBU
Fixed Pt
Load/ Store
Bin/ Hex Flt Pt
Dec Flt Pt
VBU
GR
FR

© 2012 IBM Corporation

z196 Pipeline (recap)

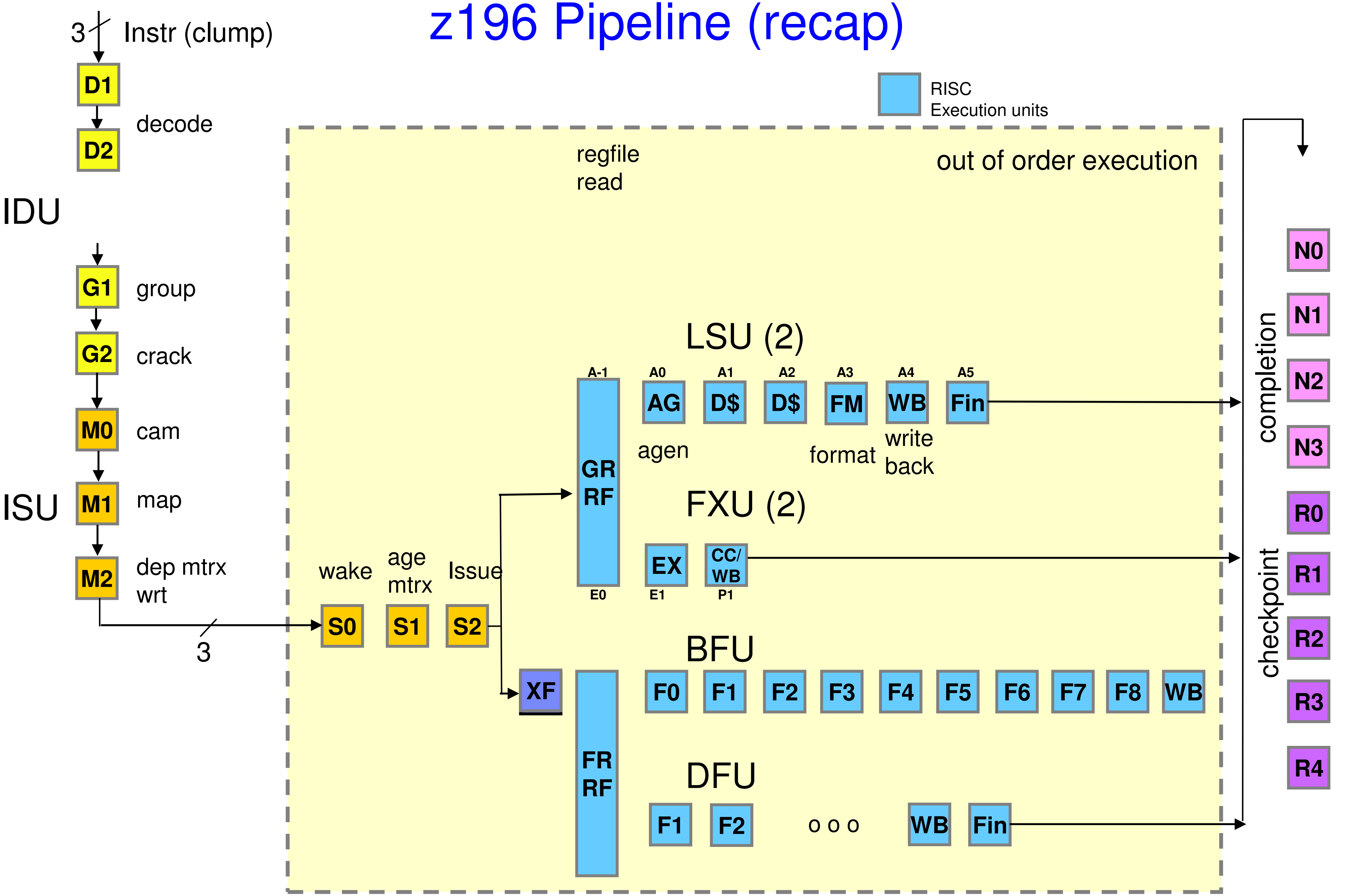
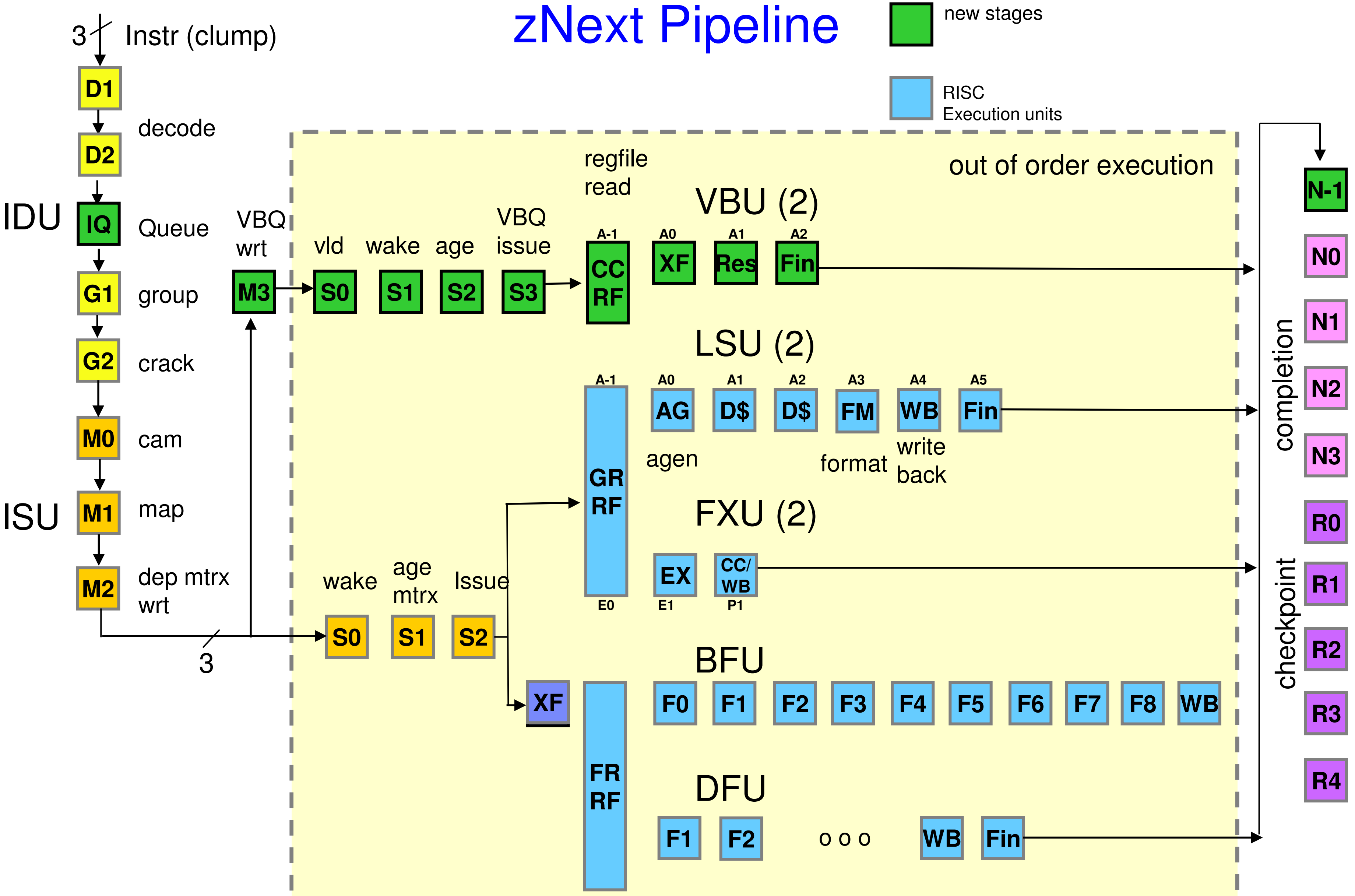


Diagram based on Brian Curran's HOTCHIP 22 presentation

© 2012 IBM Corporation

zNext Pipeline



© 2012 IBM Corporation

Feed Improvements: Accelerating Specific Functions

- Short-circuit executions
  - Common idioms executed during dispatch
  - e.g. initializing a GR with zeros

- D-Cache (SRAM design) with banking support
  - 32 banks for concurrent 2 read and 1 write operations
  - Faster cache writes reduce future load-use delays

- Dedicated Fixed-point divide engine resulting in 25-65% faster operations

- Millicode (Vertical Microcode) operations
  - Selective hardware execution

    Translate, Translate and Test, Store Clock
  - Shorter startup latency

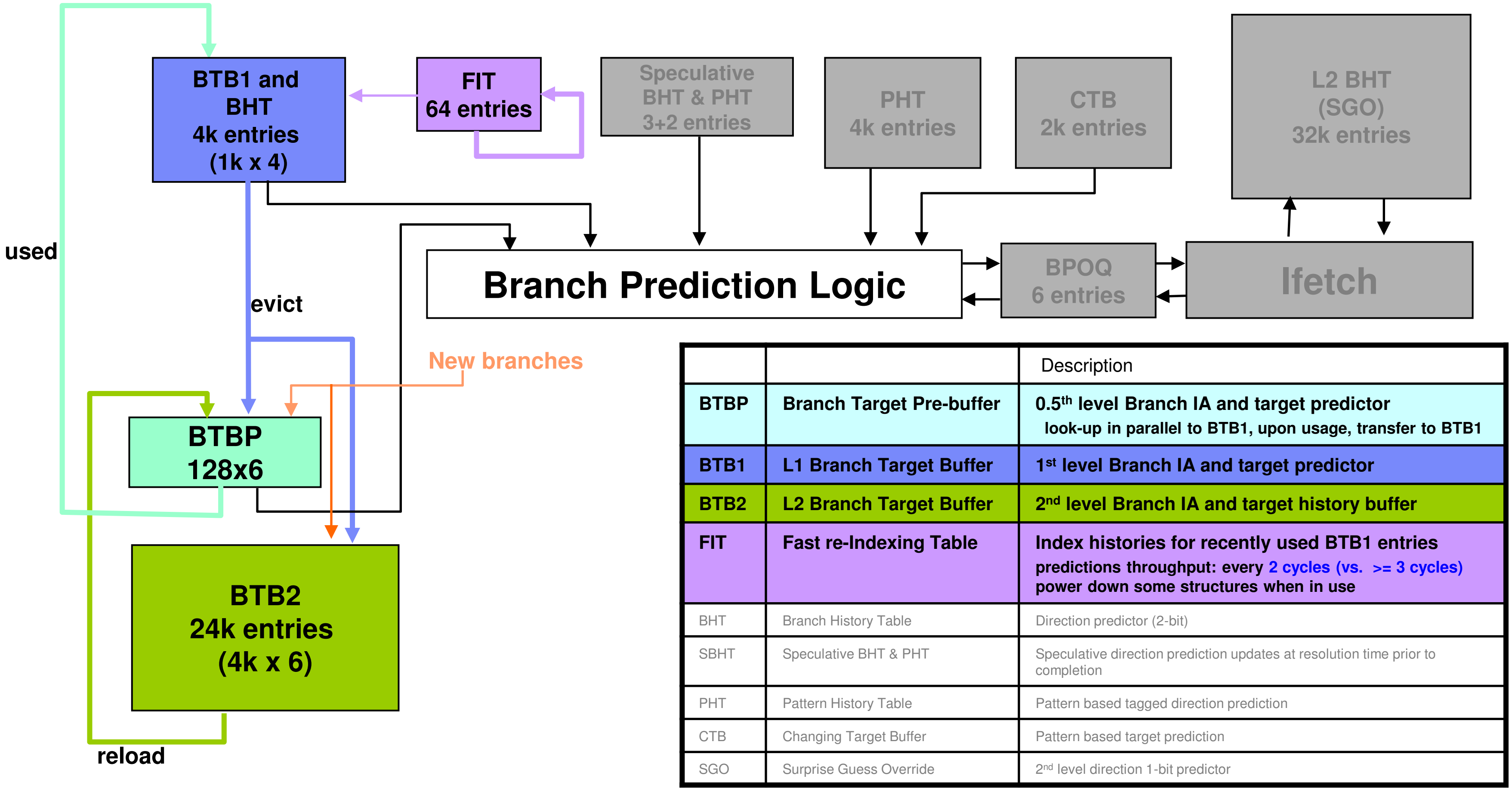
    Move Character variations, Co-Processor operations
  - Hardware assists for prefetching (target cache level & coherency state)

    Move Character Long variations
  - Dedicated hardware for Unicode conversion (UTF8 <> UTF16)

© 2012 IBM Corporation

Micro-Architecture Innovations: Branch Prediction

- Branch Prediction is essential in improving performance
  - 2nd level BTB (BTB2) for capacity (more than 3x)
  - Fast re-Indexing Table (FIT) for latency (up to 33% reduction)



		Description
BTBP	**Branch Target Pre-buffer**	**0.5th level Branch IA and target predictor** **look-up in parallel to BTB1, upon usage, transfer to BTB1**
BTB1	**L1 Branch Target Buffer**	**1st level Branch IA and target predictor**
BTB2	**L2 Branch Target Buffer**	**2nd level Branch IA and target history buffer**
FIT	**Fast re-Indexing Table**	**Index histories for recently used BTB1 entries** **predictions throughput: every 2 cycles (vs. >= 3 cycles)** **power down some structures when in use**
BHT	Branch History Table	Direction predictor (2-bit)
SBHT	Speculative BHT & PHT	Speculative direction prediction updates at resolution time prior to completion
PHT	Pattern History Table	Pattern based tagged direction prediction
CTB	Changing Target Buffer	Pattern based target prediction
SGO	Surprise Guess Override	2nd level direction 1-bit predictor

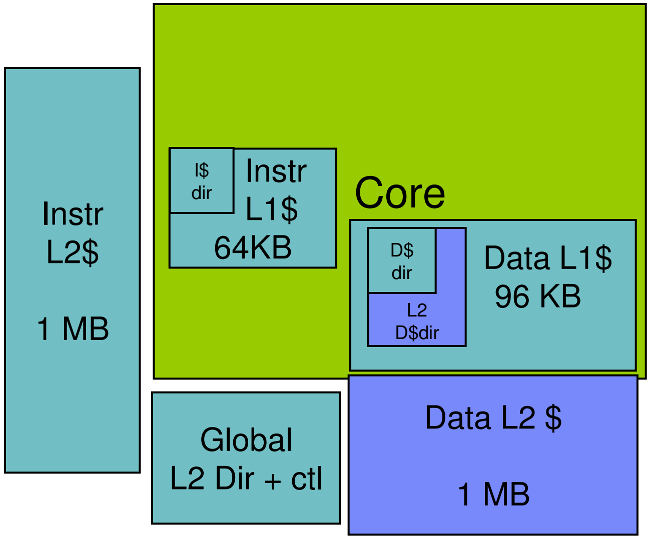
Diagram & Table based on Eric Schwarz's 2011 VAIL presentation

© 2012 IBM Corporation

Micro-Architecture Innovations: Cache Subsystem

- Split Level 2 Cache (instead of unified)
  - 1M-byte Instruction, 1M-byte Data
  - Inclusive of instruction-L1 (64 Kbyte) and data-L1 (96 Kbyte)
  - Bigger aggregate L2 with shorter latency
- Integrated data-L2 directory
  - data-L2 directory is merged into data-L1 directory
  - Logically indexed like in data-L1 directory
  - L2 Hit / miss knowledge at L1 miss time

➔ L1 miss, L2 hit latency reduced by up to 45%



Instr L2$ 1 MB

I$ dir

Instr L1$ 64KB

Core

D$ dir

L2 D$dir

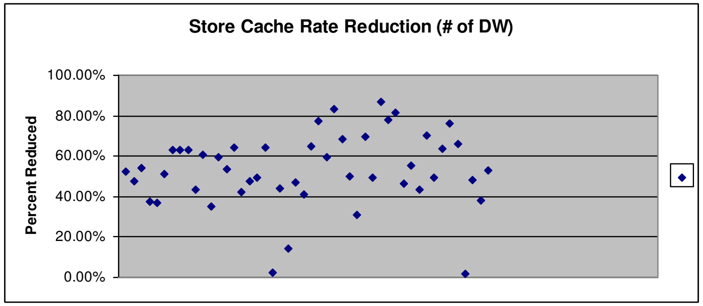
Data L1$ 96 KB

Global L2 Dir + ctl

Data L2 $ 1 MB

- Store “Gathering” Cache
  - circular queue of 64 entries of half-lines (128 bytes)
  - merges stores to same half-line post L1 updates
  - reduces pipeline usage for stores in L2 and L3
  - Hardware Transactions storage updates

➔ Store traffic to L3 typically reduced by ~50%



Modeling Data provided by Jim Mitchell @ IBM Poughkeepsie

© 2012 IBM Corporation

Targeted Architectural Extensions

- 2 Gigabyte Page support

- Decimal Floating Point Extension
  - Instructions to convert numeric data between 2 formats:
    zoned fixed-point decimal, and
    decimal floating point

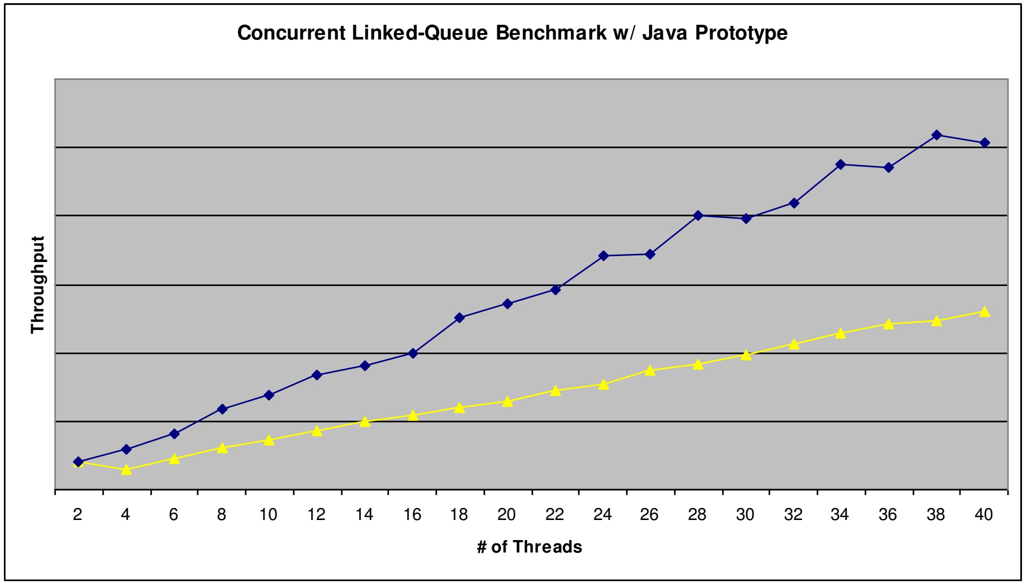
- Instruction Processing Directives
  - Branch preload instructions
    Specifies the address of a branch instruction and its target to be installed into branch prediction tables (through BTBP)

  - Data access intent instruction
    Specifies what operands of the next instruction may be further accessed for
    e.g. getting a cache line exclusive on a load for future store
    e.g. keeping access-once line at current Least-Recently-Used (LRU) position

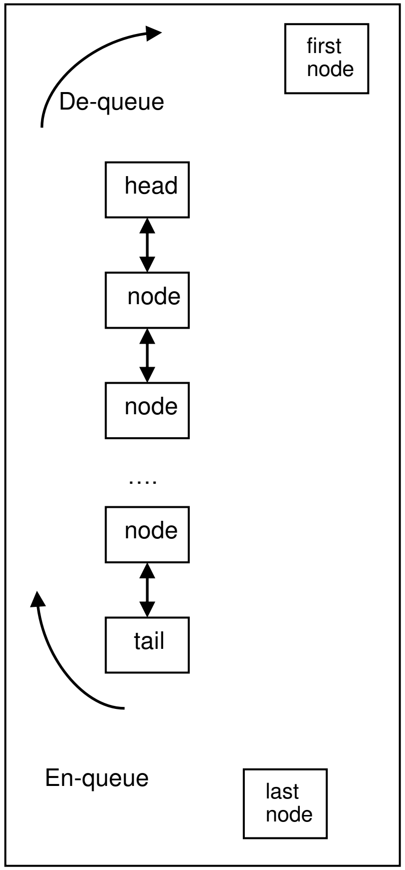
© 2012 IBM Corporation

Architectural Differentiation Extension: Transactional Execution

- General Purpose Multiprocessor Support
  - Instructions specifying start, end, and abort of a transaction
  - Pending storage updates are “shielded” from other processors until transaction completes
  - Implemented at heart of CPU (core+L1) for performance
  - Heavy focus on support for software usage and debug
  - “Constrained Transaction” with hardware auto-retries for code simplification
- Prototype benchmark with HTM
  - Showed ~2x improvements and better scalability (slope)



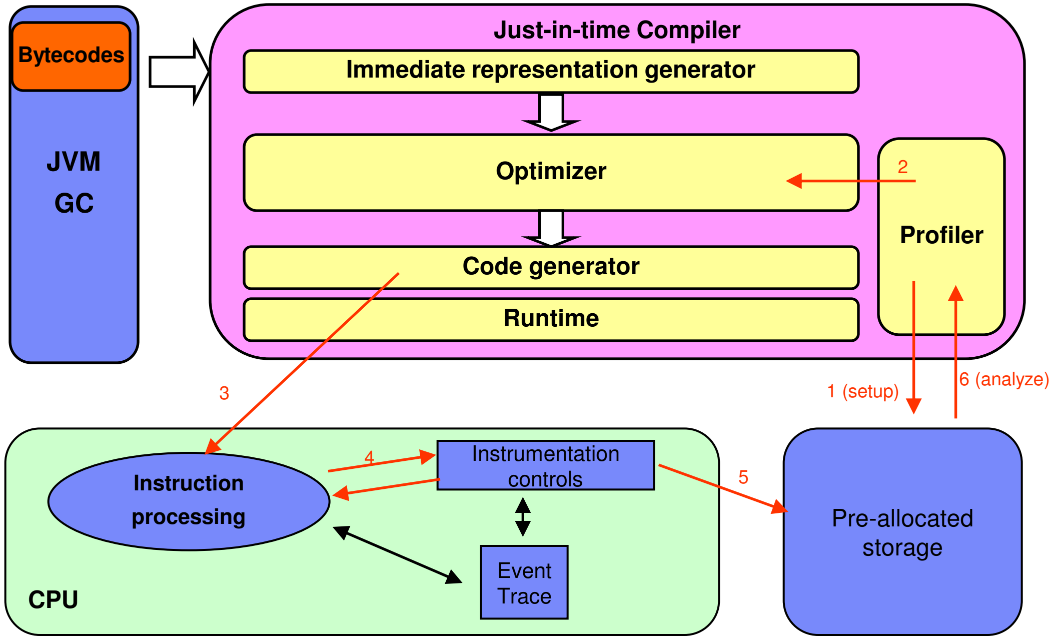
Prototype Data provided by Jerry Zheng, Marcel Mitran @ IBM Toronto

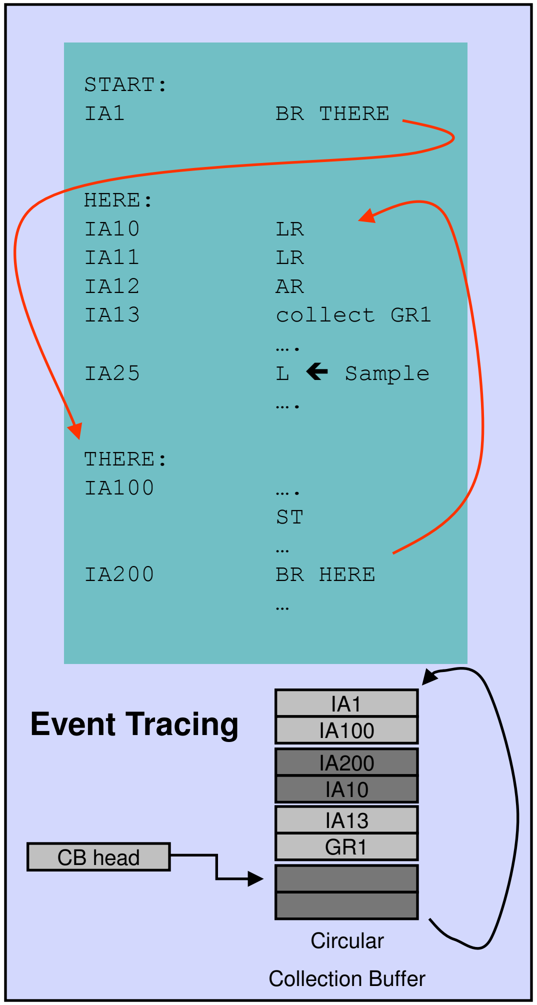


© 2012 IBM Corporation

Architectural Differentiation Extension: Runtime Instrumentation

- Low overhead profiling with hardware support
  - Instruction samples by time, count or explicit marking
- Sample reports include hard-to-get information:
  - Event traces, e.g. taken branch trace
  - "costly" events of interest, e.g. cache miss information
  - GR value profiling
- Enables better "self-tuning" opportunities



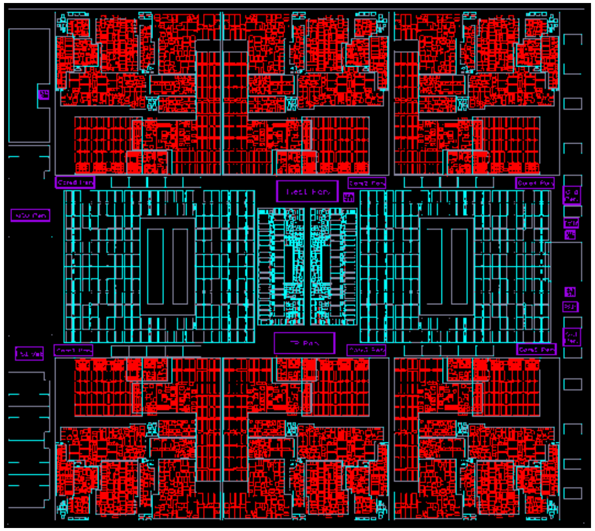


```
START:
IA1           BR THERE

HERE:
IA10          LR
IA11          LR
IA12          AR
IA13          collect GR1
              ….
IA25          L ← Sample
              ….

THERE:
IA100         ….
              ST
              …
IA200         BR HERE
              …
```

© 2012 IBM Corporation

Summary: zNext will…..

- Be used in a new family of IBM System z mainframe servers
- Sustain IBM’s mainframe leadership in computing capacity and performance without sacrificing any reliability, with
  - Up to 6 active cores per chip
  - 48M-byte shared on-chip L3 cache
  - uniquely designed low-latency private L2 cache
  - >24K target and >32k direction branch histories
  - Numerous micro-architectural enhancements*
- Provide architecture extensions*, and be the 1st general purpose microprocessor to support
  - hardware transactional memory
  - software self-directed run-time profiling
- Be amongst the fastest microprocessors @ 5.5 GHz
  - joining z196 @ continuous clock-speed of >5 GHz



* Not all features and extensions described in this presentation

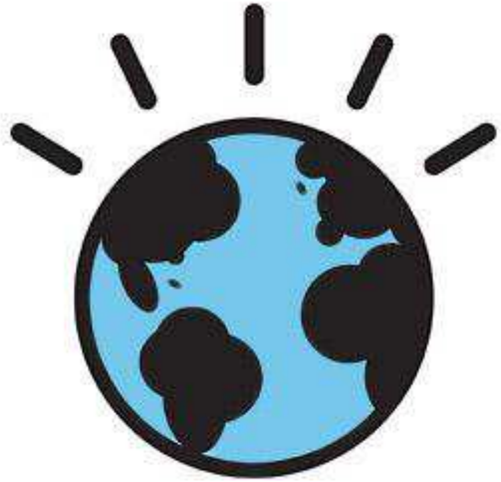
Acknowledgements

- Microarchitecture, Design and Verification Team
  - Members from
    Austin, Bangalore, Boeblingen, Haifa, Poughkeepsie, Tel Aviv, and other design labs around the world

- Architecture, Software, Performance and Research Team
  - Members from
    z/OS, z/VM, z/Linux, Compiler, JAVA, DB2, etc.

- Project Management and Technical Executives

Thank You!



© 2012 IBM Corporation

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

AIX*
BladeCenter*
CICS*
Cognos*
DataPower*
DB2*

FICON*
GDPS*
IMS
IBM*
IBM (logo)*

Parallel Sysplex*
POWER*
PR/SM
System z*
System z9*

System z10
WebSphere*
z/OS*
z/VM*
z/VSE*
zEnterprise*

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license there from.
Java and all Java-based trademarks are trademarks of Oracle Corporation in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
InfiniBand is a trademark and service mark of the InfiniBand Trade Association.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
UNIX is a registered trademark of The Open Group in the United States and other countries.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

* All other products may be trademarks or registered trademarks of their respective companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.