18-747 Lecture 14: Quantitative Case Study PPC620

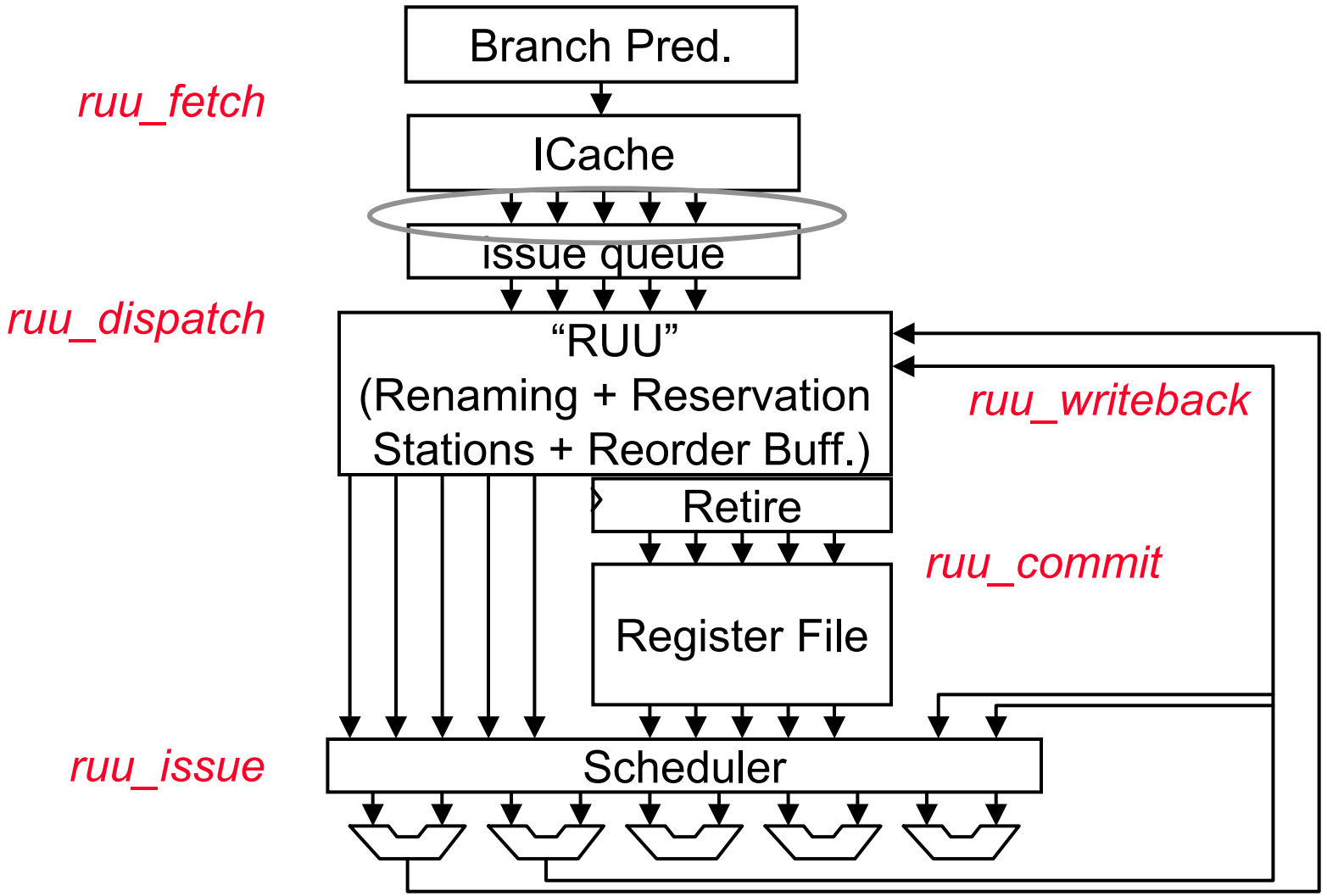
James C. Hoe
Dept of ECE, CMU
October 17, 2001

Reading Assignments: *S&L Ch3 108-130*

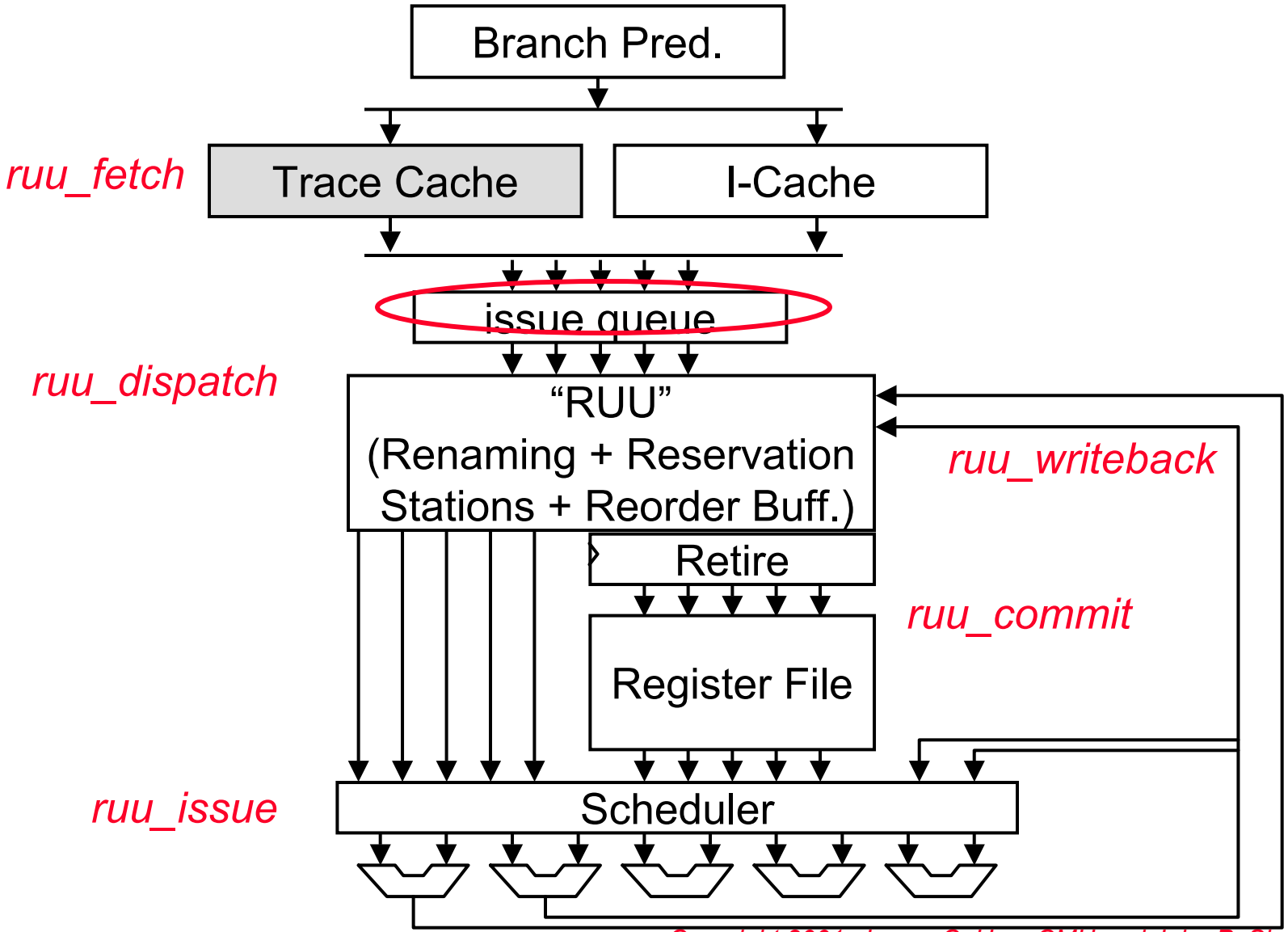
Announcements: Graded Quiz 1 will be handed back next Wednesday
No lecture on Monday
Yale Patt, Noon - 1 pm, Newell Simon Hall 3305

Handouts: *Handout10: Project 2*
Special Take-Home Quiz

Project 2: Sim-Outorder



Project 2: Basic Trace Cache



Special Properties at the "Interface"

- 30,000-feet View of Prediction and Speculation
  - Branch prediction can generate next-PC at random
  - The executing stages will catch any mistakes and force restart at the correct PC
- 3 feet Away
  - BTB never makes a mistake on a non-control-flow instruction

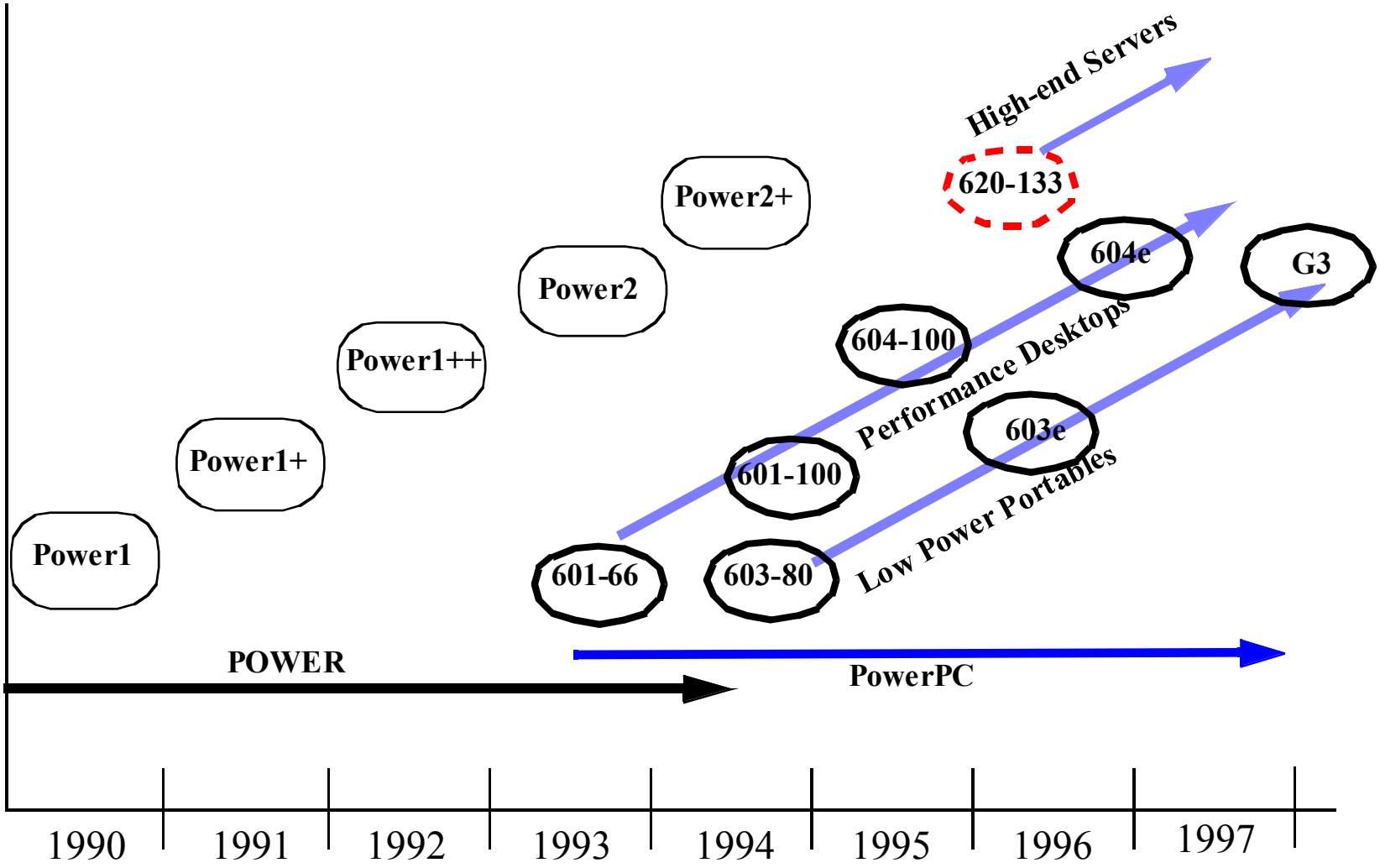
    ⇒ *Hardware only have to check right after a control-flow instruction*
  - Your trace cache must feed correct sequences within a basic block *(i.e. next-PC=PC+4)*

    ⇒ *Any "reasonably" collected trace should*
- Down in the dirt
  - How long should traces be?
  - Collect at the front (speculative traces) or back (commit traces)?
  - What to do when traces collide?
  - How to reduce redundant storage?
  - How to find the best trace?

The PowerPC Architecture

- IBM, Motorola, and Apple Alliance
- Based on the IBM POWER Architecture
  - Facilitate parallel execution
  - Scale well with advancing technology
- Instruction Set Architecture
  - 32 integer registers (GPRs)
  - 32 floating-point registers (FPRs)
  - 1 condition register (CR)
  - 1 count register (CTR)
  - 1 link register (LR)
  - 1 integer exception register (XER)
  - 1 floating point status and control register (FPSCR)
  - Load/store architecture; Typical RISC instructions

The PowerPC Roadmap

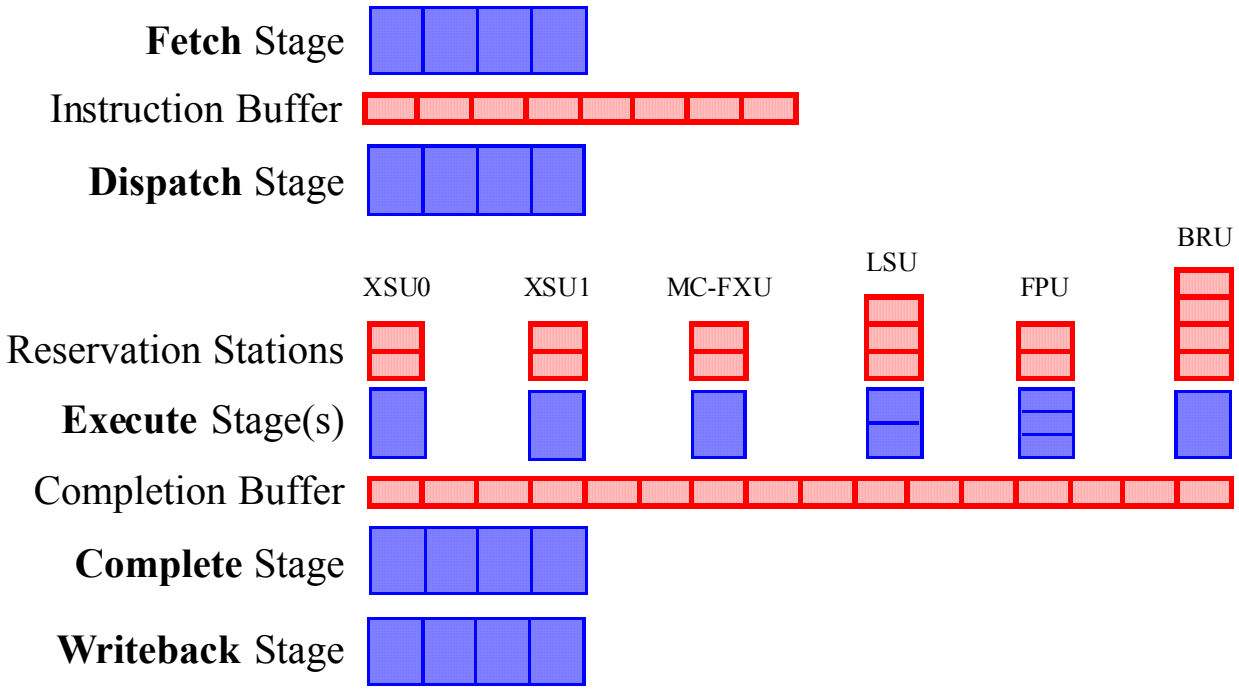


The PowerPC 620 Implementation

- Out-of-order execution
- Distributed reservation stations
- Register renaming
- Dynamic branch prediction
- Speculative execution
- Weak memory ordering
- Support for precise exception

620 Instruction Pipeline

- 5 major pipeline stages
- 3 sets of buffers between stages
- 4-wide superscalar
- 6 execution units

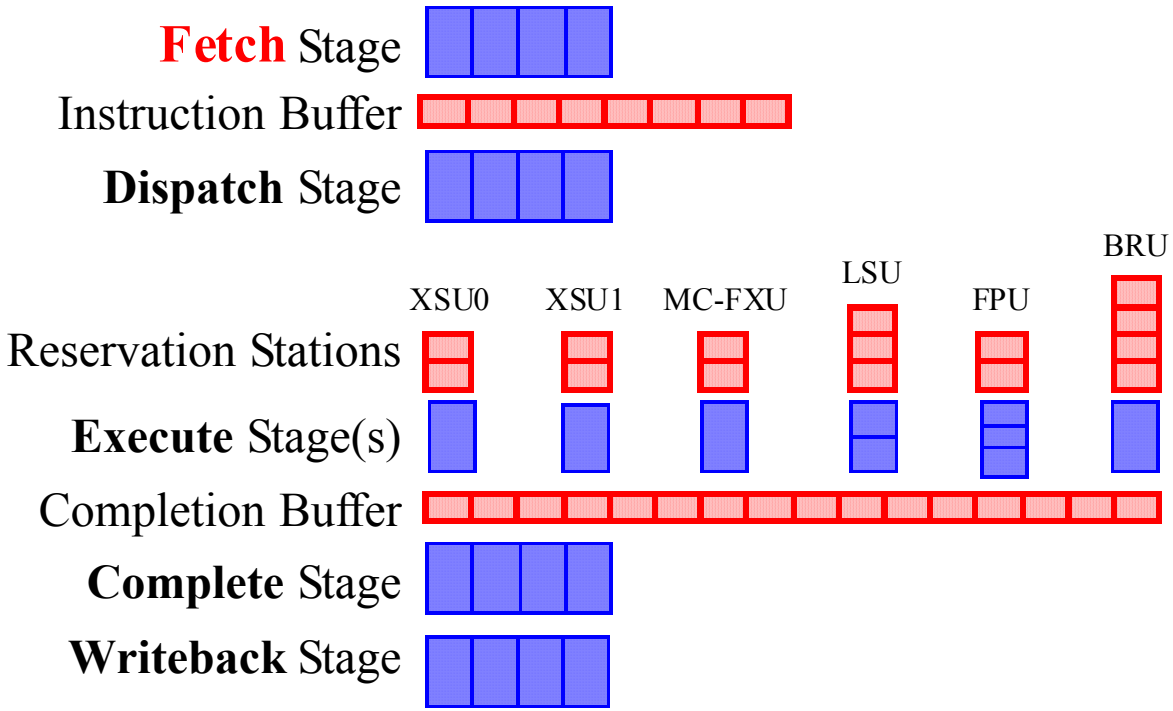


Cache Configurations

- L1 I-cache
  - 32 kilobytes
  - 8-way associative
- L1 D-cache
  - 32 kilobytes
  - 8-way associative
  - 2-way interleaving
  - 2 ports: 1 for load, 1 for store
  - Nonblocking cache
  - Allow weak ordering of memory accesses
- Dedicated back-door interface to a unified L2 cache
  - On chip tag banks, off-chip data banks
  - ~10 cycles latency

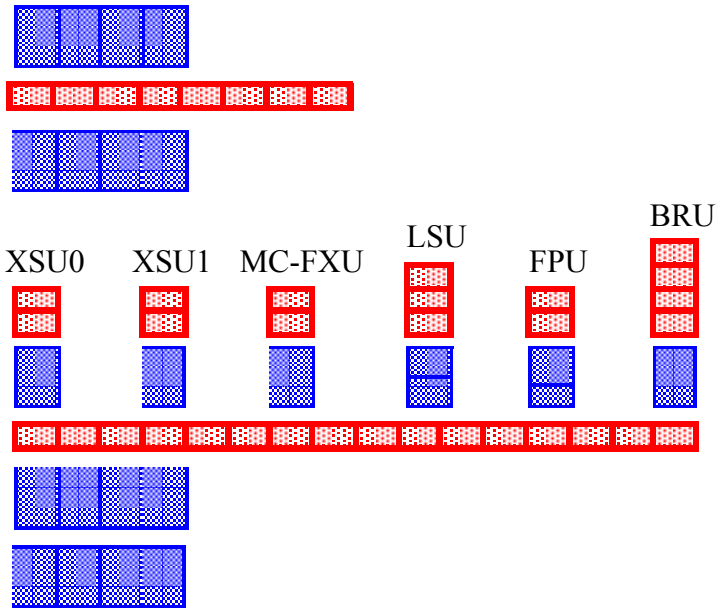
Instruction Fetching

- Access the instruction cache
- Fetch up to 4 instructions per cycle
- Fetch cannot cross cache lines



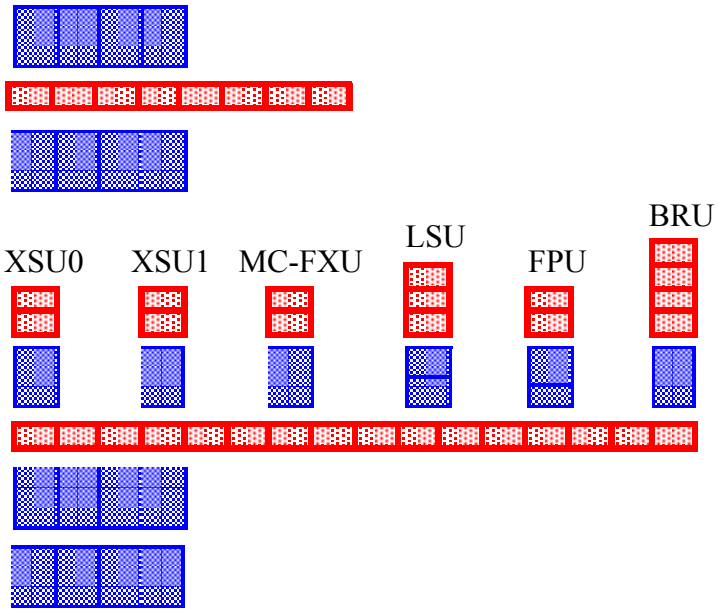
Instruction Dispatching

- Check instructions for dispatch from the instruction buffer to the reservation stations
  - If ready for dispatch, allocate a reservation station entry, a completion buffer entry, and a rename buffer entry for the destination operand
  - In-order dispatch, up to four instructions per cycle
- Sources of dispatch stall
  - Reservation station saturation
  - Another dispatched to the same unit
  - Completion buffer saturation
  - Register read port saturation
  - Rename buffer saturation
  - Branch wait for mtspr
  - Serialization constraint



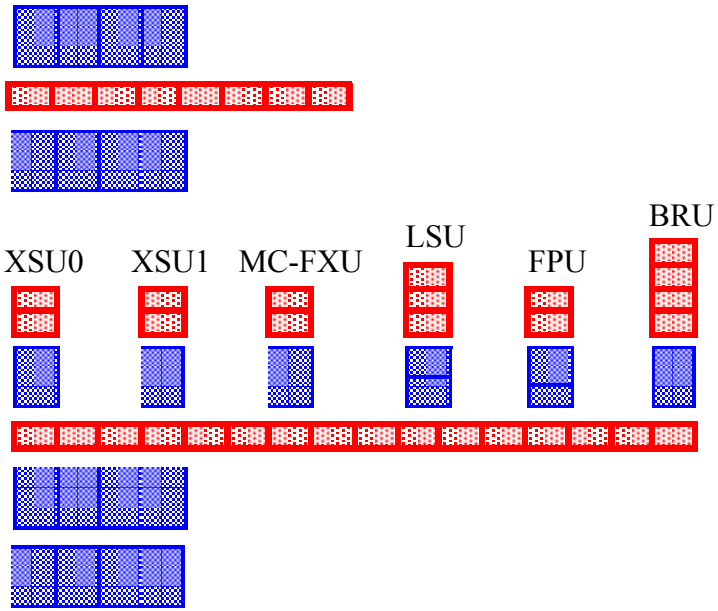
Instruction Execution

- Check instructions in the reservation station for issue into the execution unit
  - Check for operand ready, possibly via forwarding
  - If ready for issue, execute the instruction
  - Out-of-order issue
- Sources of issue stall
  - Out of order disallowed in some units
  - Waiting for source operand
  - Waiting for execution unit
  - Serialization constraint



Instruction Completion and Writeback

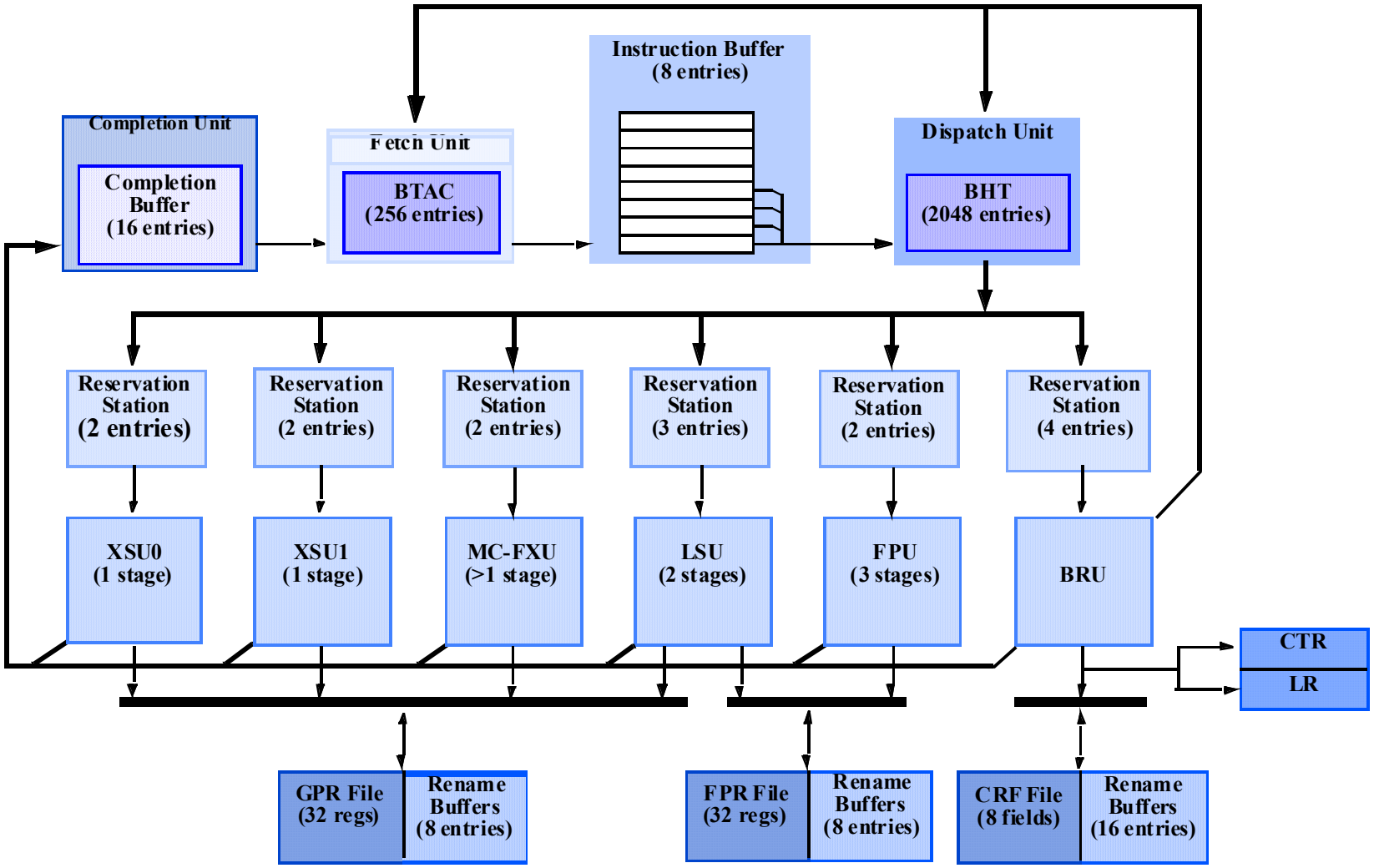
- When instructions finish execution, remove the instructions from the completion buffer according to program order
  - Instructions in the completion buffer: in-flight instructions
  - Completion buffer supports precise exception
- Commit the results of the instructions to the architected register files and memory
  - Most instructions copy results from rename registers to architected registers.
  - Store instructions access cache



Branch Prediction

- Fetch stage
  - Access the Branch Target Address Cache (256 entries)
  - If miss in BTAC, continue fetching at the next sequential address
  - If hit, fetch instructions at the target address
- Dispatch stage
  - Resolve unconditional branch
  - If condition ready, resolve conditional branch
  - If not ready, predict branch direction using the BHT (2048 entries)
  - Two history bits used as a hysteresis counter
  - If prediction result different from BTAC's, then use BHT's prediction
  - After branch is resolved, update the BTAC
- Speculation support for up to 4 predicted branches

620 Block Diagram



Experimental Framework

- Current trends in design methods
  - Quantitative approach
  - Benchmarking of real programs
  - Use of software tools to collect performance data
- Visualization-based Microarchitecture Workbench
  - Machine specification
  - Simulator compilation
  - Trace-driven simulation

Benchmarks

Benchmark Type	Benchmarks	Dynamic Instructions
Integer	*compress*	6,884,247
	eqntott	3,147,233
	espresso	4,615,085
	li	3,376,415
Floating Point	*alvinn*	4,861,138
	hydro2d	4,114,602
	tomcatv	6,858,619

Instruction Mix

Instruction Mix		Integer Benchmarks (SPECint 92)				Floating-Point Benchmarks (SPECfp 92)		
		compress	*eqntott*	*espresso*	*li*	*alvinn*	*hydro2d*	*tomcatv*
Integer	**Arith. (single cycle)**	42.73%	48.79%	48.30%	29.54%	37.50%	26.25%	19.93%
	Arith. (multi-cycle)	0.89%	1.26%	1.25%	5.14%	0.29%	1.19%	0.05%
	Load	25.39%	23.21%	24.34%	28.48%	0.25%	0.46%	0.31%
	Store	16.49%	6.26%	8.29%	18.60%	0.20%	0.19%	0.29%
Floating-Point	**Arith. (pipelined)**	0.00%	0.00%	0.00%	0.00%	12.27%	26.99%	37.82%
	Arith. (non-pipelined)	0.00%	0.00%	0.00%	0.00%	0.08%	1.87%	0.70%
	Load	0.00%	0.00%	0.00%	0.01%	26.85%	22.53%	27.84%
	Store	0.00%	0.00%	0.00%	0.01%	12.02%	7.74%	9.09%
Branch	**Unconditional**	1.90%	1.87%	1.52%	3.26%	0.15%	0.10%	0.01%
	Conditional	12.15%	17.43%	15.26%	12.01%	10.37%	12.50%	3.92%
	Cond. to Count Reg.	0.00%	0.44%	0.10%	0.39%	0.00%	0.16%	0.05%
	Cond. to Link Reg.	4.44%	0.74%	0.94%	2.55%	0.03%	0.01%	0.00%

Performance Evaluation Data

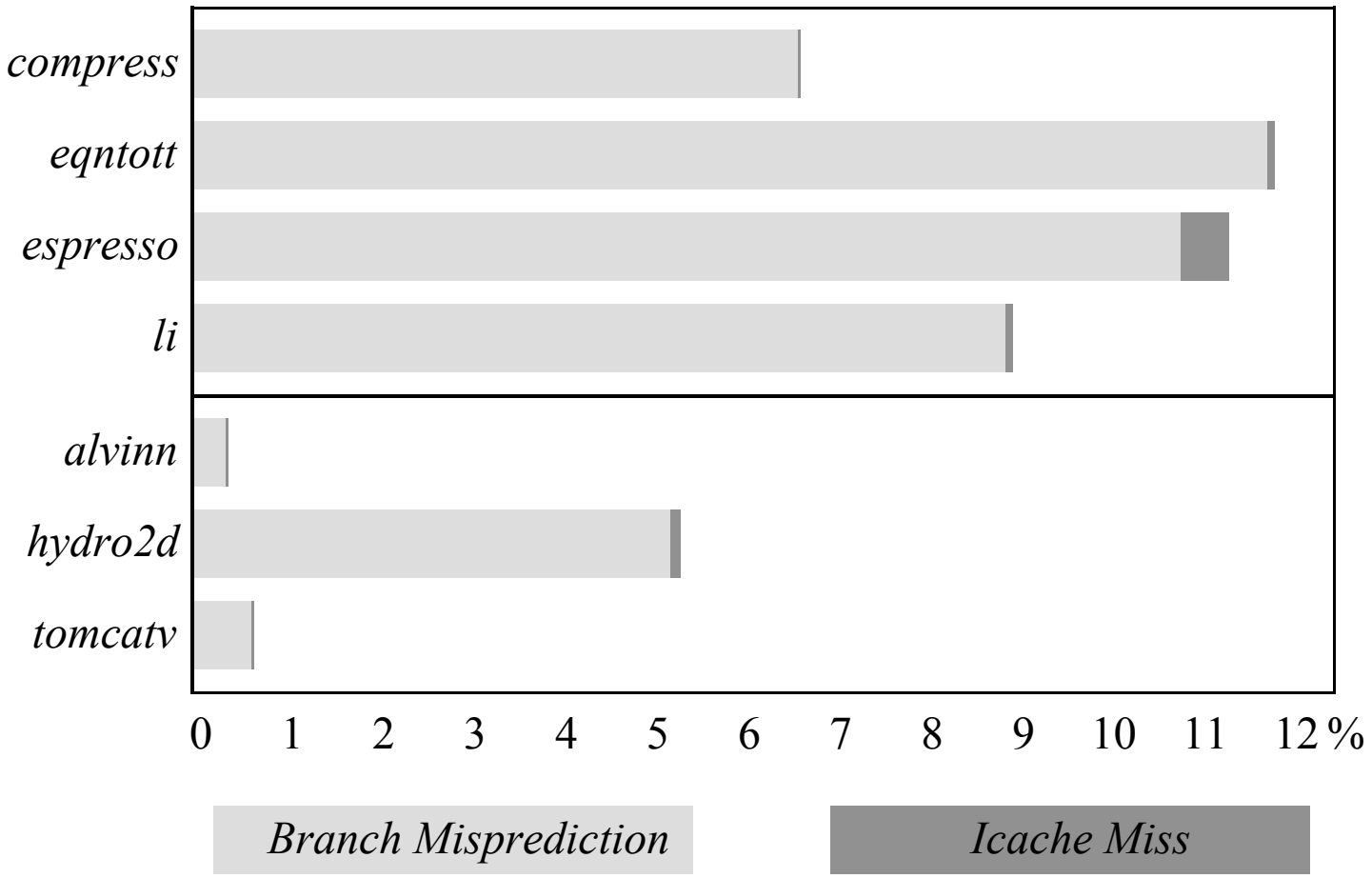
- Overall Sustained IPC
  - Integer benchmarks: 0.99 to 1.44
  - Floating-point benchmarks: 0.96 to 1.77

Benchmarks	Dynamic Instructions	Execution Cycles	IPC
compress	6,884,247	6,062,494	1.14
eqntott	3,147,233	2,188,331	1.44
espresso	4,615,085	3,412,653	1.35
li	3,376,415	3,399,293	0.99
alvinn	4,861,138	2,744,098	1.77
hydro2d	4,114,602	4,293,230	0.96
tomcatv	6,858,619	6,494,912	1.06

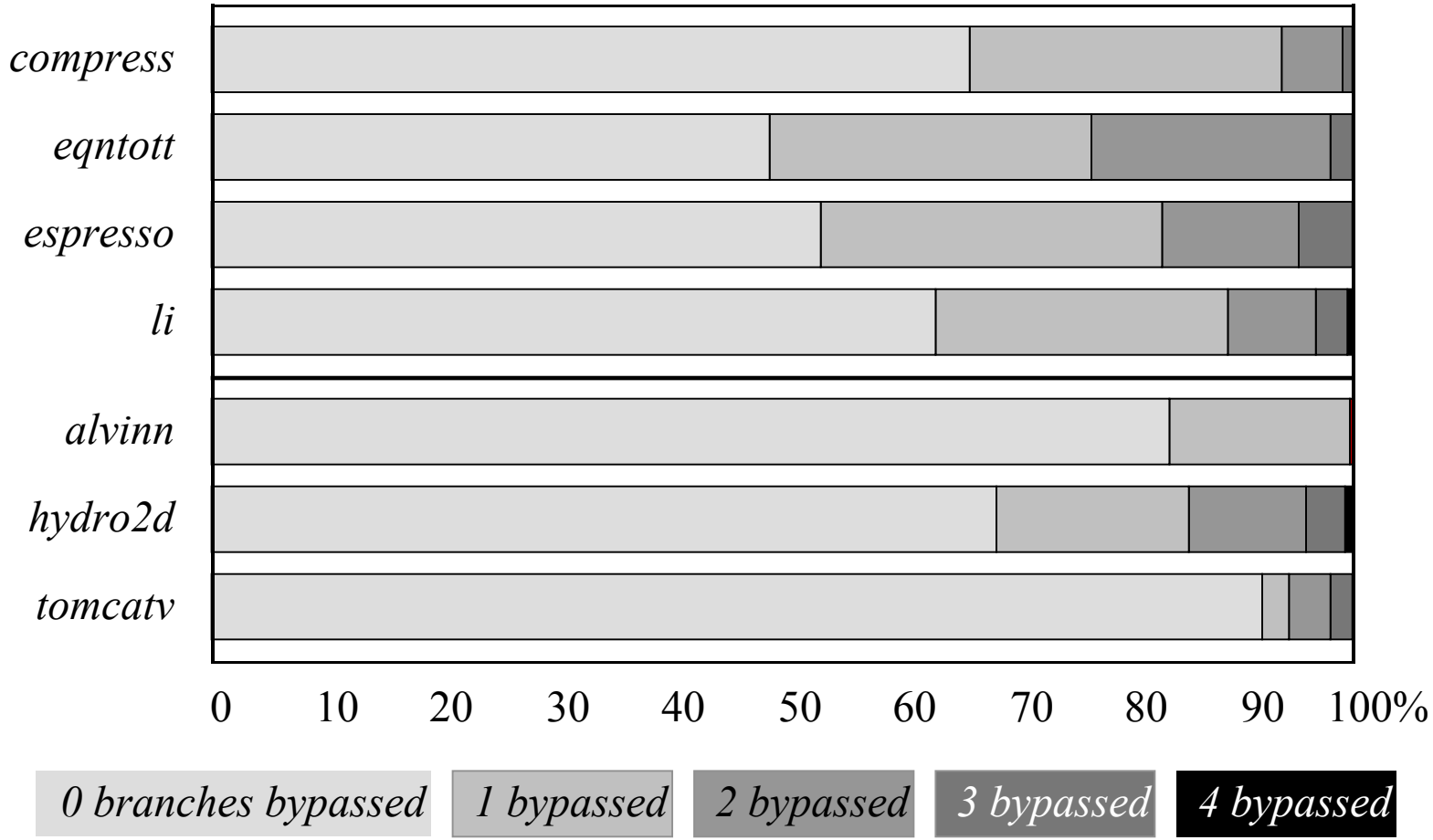
Branch Prediction Data

Branch Processing		*compress*	*eqntott*	*espresso*	*li*	*alvinn*	*hydro2d*	*tomcatv*
Branch Resolution	**Not Taken**	40.35%	31.84%	40.05%	33.09%	6.38%	17.51%	6.12%
	Taken	59.65%	68.16%	59.95%	66.91%	93.62%	82.49%	93.88%
BTAC Prediction	**Correct**	84.10%	82.64%	81.99%	74.70%	94.49%	88.31%	93.31%
	Incorrect	15.90%	17.36%	18.01%	25.30%	5.51%	11.69%	6.69%
BHT Prediction	**Resolved**	19.71%	18.30%	17.09%	28.83%	17.49%	26.18%	45.39%
	Correct	68.86%	72.16%	72.27%	62.45%	81.58%	68.00%	52.56%
	Incorrect	11.43%	9.54%	10.64%	8.72%	0.92%	5.82%	2.05%
BTAC Incor. & BHT Correct		0.01%	0.79%	1.13%	7.78%	0.07%	0.19%	0.00%
BTAC Correct & BHT Incor.		0.00%	0.12%	0.37%	0.26%	0.00%	0.08%	0.00%
Overall Br. Pred. Accuracy		88.57%	90.46%	89.36%	91.28%	99.07%	94.18%	97.95%

Zero-Bandwidth Fetch Cycles



Degree of Speculation



Average Buffer Usage

Buffer Usage	*compress*	*eqntott*	*espresso*	*li*	*alvinn*	*hydro2d*	*tomcatv*
XSU0 RS Entries (2)	0.37	0.66	0.68	0.36	0.48	0.23	0.11
XSU1 RS Entries (2)	0.42	0.51	0.65	0.32	0.24	0.17	0.10
MC-FXU RS Entries (2)	0.04	0.07	0.09	0.28	0.01	0.10	0.00
FPU RS Entries (2)	0.00	0.00	0.00	0.00	0.70	1.04	0.89
LSU RS Entries (3)	1.69	1.36	1.60	1.73	2.26	0.98	1.23
BRU RS Entries (4)	0.45	0.84	0.75	0.59	0.19	0.54	0.17
GPR Rename Buffers (8)	2.73	3.70	3.25	2.77	3.79	1.83	1.97
FPR Rename Buffers (8)	0.00	0.00	0.00	0.00	5.03	2.85	3.23
CR Rename Buffers (16)	1.25	1.32	1.19	0.98	1.27	1.20	0.42
Completion Buffers (16)	10.75	8.83	8.75	9.87	13.91	10.10	11.16

Dispatch Stalls

Sources of Dispatch Stalls	*compress*	*eqntott*	*espresso*	*li*	*alvinn*	*hydro2d*	*tomcatv*
Serialization	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Move to special reg. const.	0.00%	4.49%	0.94%	3.44%	0.00%	0.95%	0.08%
Read port saturation	0.26%	0.00%	0.02%	0.00%	0.32%	2.23%	6.73%
Reservation station sat.	36.07%	22.36%	31.50%	34.40%	22.81%	42.70%	36.51%
Rename buffer saturation	24.06%	7.60%	13.93%	17.26%	1.36%	36.98%	34.13%
Completion buffer sat.	5.54%	3.64%	2.02%	4.27%	21.12%	7.80%	9.03%
Another to same unit	9.72%	20.51%	18.31%	10.57%	24.30%	12.01%	7.17%
No dispatch stalls	24.35%	41.40%	33.28%	30.06%	30.09%	17.33%	6.35%

Issue Stalls

Sources of Dispatch Stalls	*compress*	*eqntott*	*espresso*	*li*	*alvinn*	*hydro2d*	*tomcatv*
Out of order disallowed	0.00%	0.00%	0.00%	0.00%	0.72%	11.03%	1.53%
Serialization	1.69%	1.81%	3.21%	10.81%	0.03%	4.47%	0.01%
Waiting for Source	21.97%	29.30%	37.79%	32.03%	17.74%	22.71%	3.52%
Waiting for execution unit	13.67%	3.28%	7.06%	11.01%	2.81%	1.50%	1.30%
No issue stalls	62.67%	65.61%	51.94%	46.15%	78.70%	60.29%	93.64%

Not including branch unit

Average Execution Latency

Function Units	*compress*	*eqntott*	*espresso*	*li*	*alvinn*	*hydro2d*	*tomcatv*
XSU0 (1 stage)	1.53	1.62	1.89	2.28	1.05	1.48	1.01
XSU1 (1 stage)	1.72	1.73	2.23	2.39	1.13	1.78	1.03
MC-FXU (>1 stage)	4.35	4.82	6.18	5.64	3.48	9.61	1.64
FPU (3 stages)	----*	----*	----*	----*	5.29	6.74	4.45
LSU (2 stages)	3.56	2.35	2.87	3.22	2.39	2.92	2.75
BRU	2.71	2.86	3.11	3.28	1.04	4.42	4.14

Very few instructions are executed in this unit for these benchmarks

Cache Effects

Cache Effects	*compress*	*eqntott*	*espresso*	*li*	*alvinn*	*hydro2d*	*tomcatv*
Loads/stores with cache hit	94.17%	99.57%	99.92%	99.74%	99.99%	94.58%	96.24%
Loads that bypass a missed load	8.45%	0.53%	0.11%	0.14%	0.01%	4.82%	5.45%
Loads that bypass a pending store	58.85%	21.05%	27.17%	48.49%	98.33%	58.26%	43.23%
Load that aliased with a pending store	0.00%	0.31%	0.77%	2.59%	0.27%	0.21%	0.29%
Avg. number of pending stores per cycle	1.96	0.83	0.97	2.11	1.30	1.01	1.38