

COMPAQ

The Alpha 21264 Microprocessor: Out-of-Order Execution at 600 MHz

R. E. Kessler
Compaq Computer Corporation
Shrewsbury, MA

REK August 1998

1

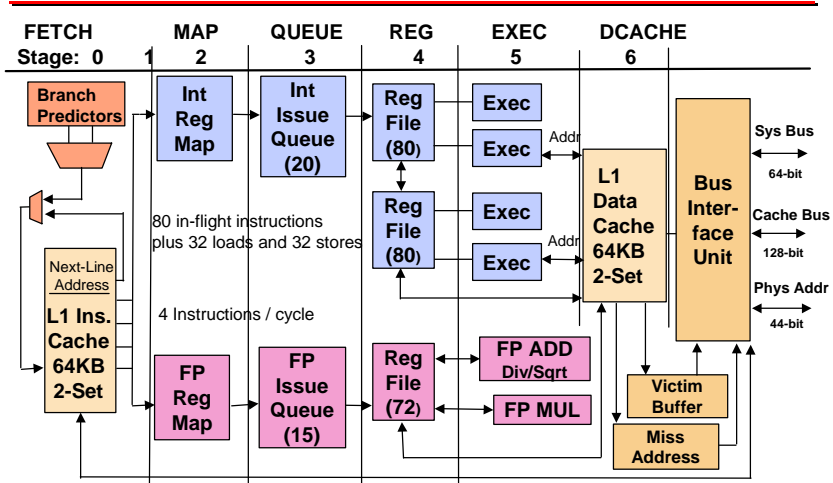
Some Highlights

- **Continued Alpha performance leadership**
 - 600 MHz operation in 0.35u CMOS6, 6 metal layers, 2.2V
 - 15 Million transistors, 3.1 cm², 587 pin PGA
 - Specint95 of 30+ and Specfp95 of 50+
 - Out-of-order and speculative execution
 - 4-way integer issue
 - 2-way floating-point issue
 - Sophisticated tournament branch prediction
 - High-bandwidth memory system (1+ GB/sec)

REK August 1998

2

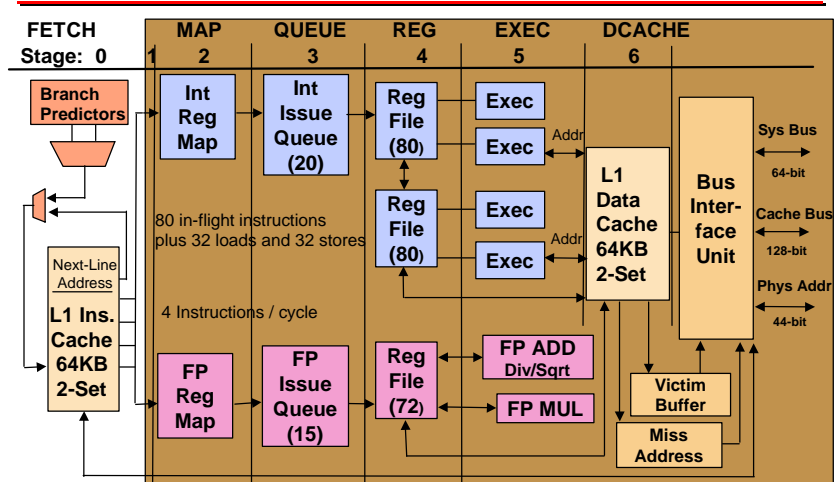
Alpha 21264: Block Diagram



REK August 1998

3

Alpha 21264: Block Diagram



REK August 1998

4

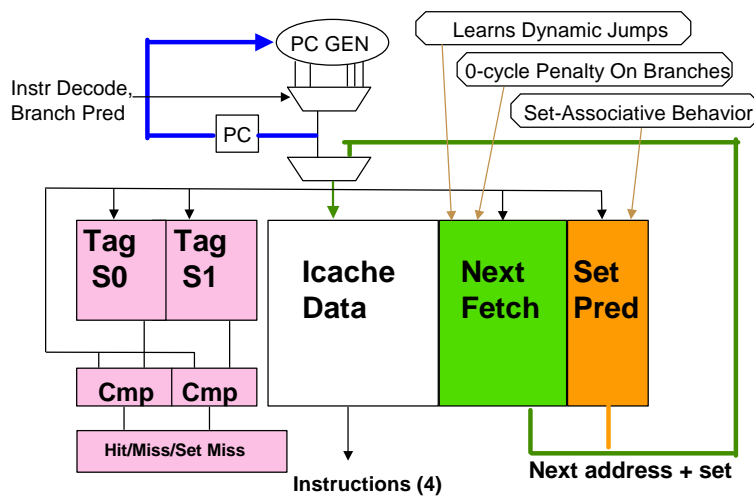
21264 Instruction Fetch Bandwidth Enablers

- The 64 KB two-way associative instruction cache supplies four instructions every cycle
- The *next-fetch* and *set predictors* provide the fast cache access times of a direct-mapped cache and eliminate bubbles in non-sequential control flows
- The instruction fetcher speculates through up to 20 branch predictions to supply a continuous stream of instructions
- The tournament branch predictor dynamically selects between *Local* and *Global* history to minimize mispredicts

REK August 1998

5

Instruction Stream Improvements

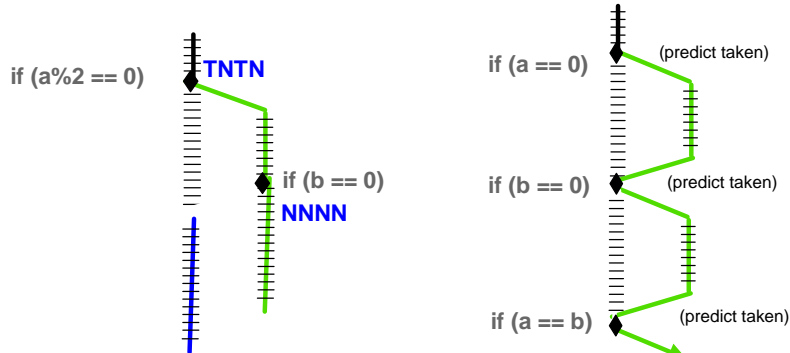


REK August 1998

6

Fetch Stage: Branch Prediction

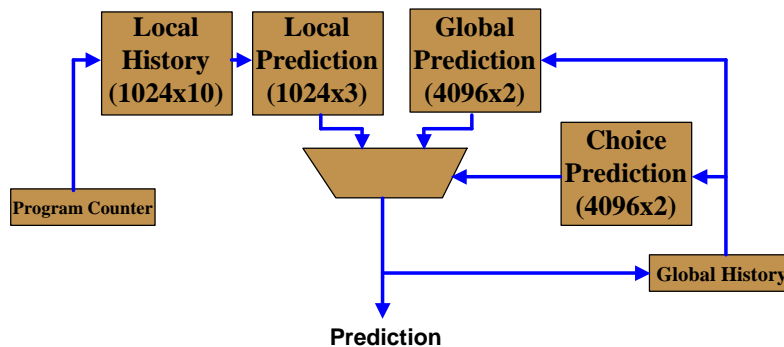
- Some branch directions can be predicted based on their past behavior: *Local Correlation*
- Others can be predicted based on how the program arrived at the branch: *Global Correlation*



REK August 1998

7

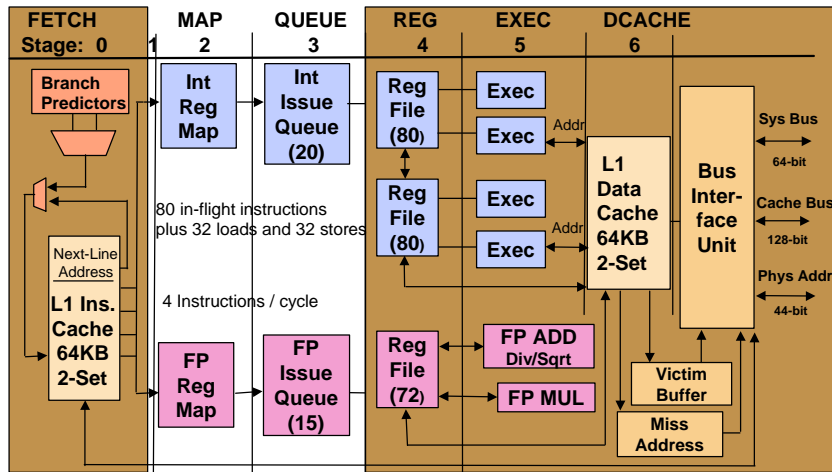
Tournament Branch Prediction



REK August 1998

8

Alpha 21264: Block Diagram



REK August 1998

9

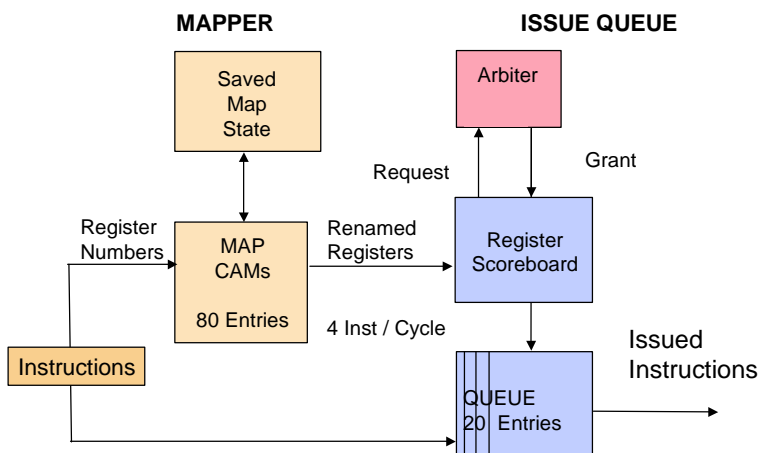
Mapper and Queue Stages

- **Mapper:**
 - Rename 4 instructions per cycle (8 source / 4 dest)
 - 80 integer + 72 floating-point physical registers
- **Queue Stage:**
 - Integer: 20 entries / Quad-Issue
 - Floating Point: 15 entries / Dual-Issue
 - Instructions issued out-of-order when data ready
 - Prioritized from oldest to youngest each cycle
 - Instructions leave the queues after they issue
 - The queue collapses every cycle as needed

REK August 1998

10

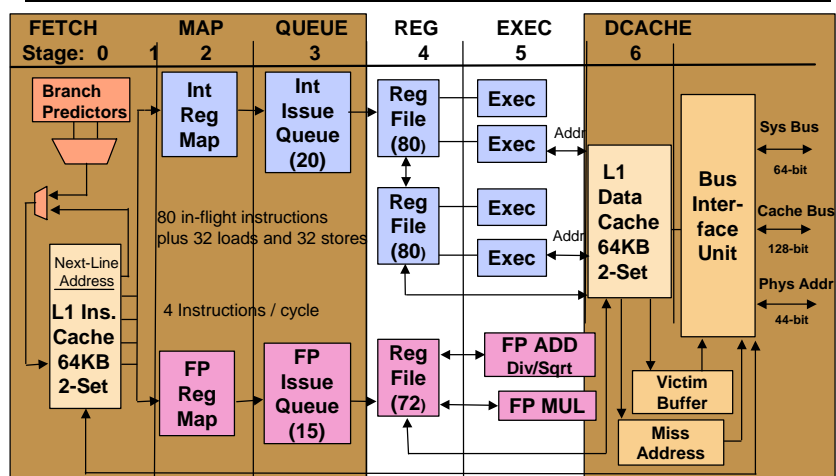
Map and Queue Stages



REK August 1998

11

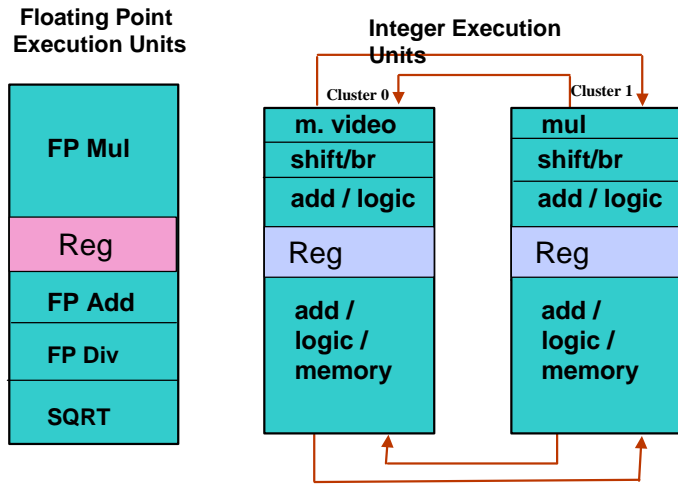
Alpha 21264: Block Diagram



REK August 1998

12

Register and Execute Stages

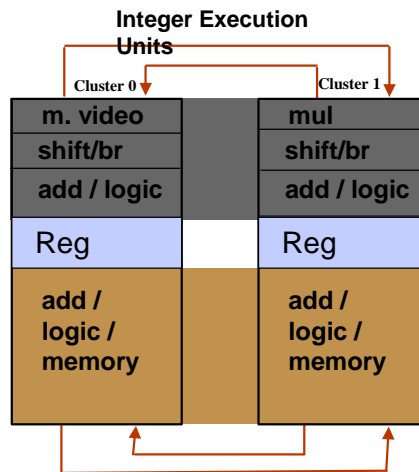


REK August 1998

13

Integer Cross-Cluster Instruction Scheduling and Execution

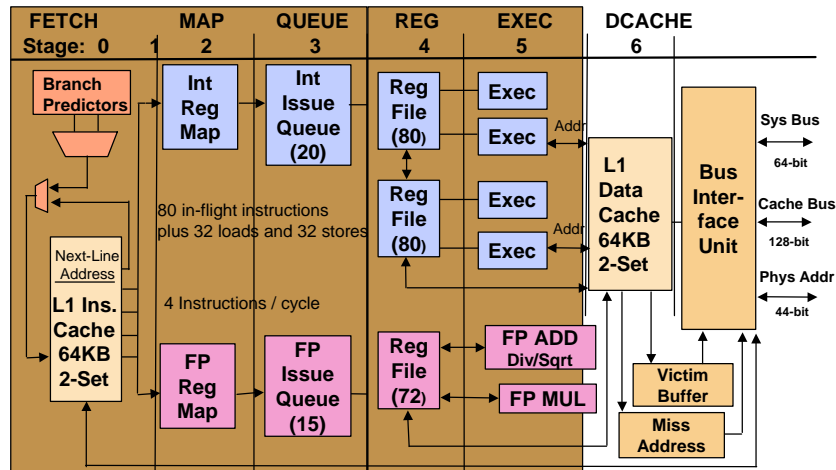
- Instructions are statically pre-slotted to the upper or lower execution pipes
- The issue queue dynamically selects between the left and right clusters
- This has most of the performance of 4-way with the simplicity of 2-way issue



REK August 1998

14

Alpha 21264: Block Diagram



REK August 1998

15

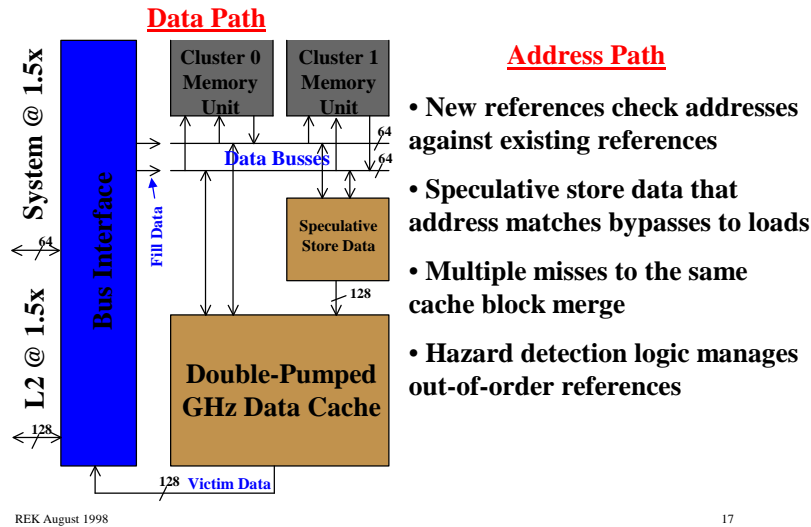
21264 On-Chip Memory System Features

- **Two loads/stores per cycle**
Any combination
- **64 KB two-way associative L1 data cache (9.6 GB/sec)**
Phase-pipelined at > 1 GHz (no bank conflicts!)
3 cycle latency (issue to issue of consumer) with hit prediction
- **Out-of-order and speculative execution**
Minimizes effective memory latency
- **32 outstanding loads and 32 outstanding stores**
Maximizes memory system parallelism
- **Speculative stores forward data to subsequent loads**

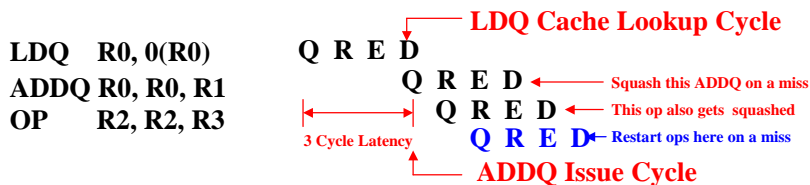
REK August 1998

16

Memory System (Continued)

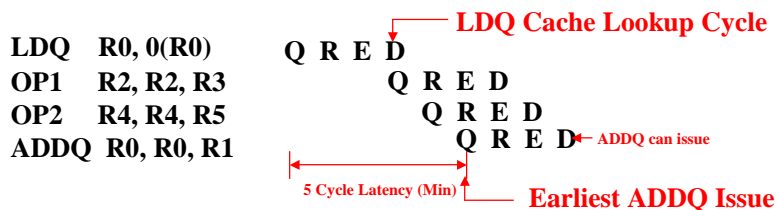


Low-latency Speculative Issue of Integer Load Data Consumers (Predict Hit)



- **When predicting a load hit:**
 - The ADDQ issues (speculatively) after 3 cycles
 - Best performance if the load actually hits (matching the prediction)
 - The ADDQ issues before the load hit/miss calculation is known
- **If the LDQ misses when predicted to hit:**
 - Squash two cycles (replay the ADDQ and its consumers)
 - Force a “mini-replay” (direct from the issue queue)

Low-latency Speculative Issue of Integer Load Data Consumers (Predict Miss)



- **When predicting a load miss:**
 - The minimum load latency is 5 cycles (more on a miss)
 - There are no squashes
 - Best performance if the load actually misses (as predicted)
- **The hit/miss predictor:**
 - MSB of 4-bit counter (hits increment by 1, misses decrement by 2)

REK August 1998

19

Dynamic Hazard Avoidance (Before Marking)

Program order (Assume R28 == R29):

LDQ R0, 64(R28)
 STQ R0, 0(R28) ← Store followed by load to the
 LDQ R1, 0(R29) ← same memory address!

execution order:

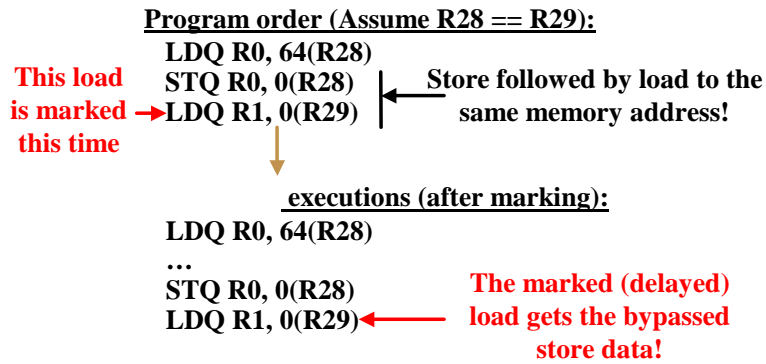
LDQ R0, 64(R28)
 LDQ R1, 0(R29) ←
 ...
 STQ R0, 0(R28)

This (re-ordered) load got the wrong data value!

REK August 1998

20

Dynamic Hazard Avoidance (After Marking/Training)



REK August 1998

21

New Memory Prefetches

● Software-directed prefetches

- Prefetch
- Prefetch w/ modify intent
- Prefetch, evict it next
- Evict Block (eject data from the cache)
- Write hint
 - allocate block with no data read
 - useful for full cache block writes

REK August 1998

22

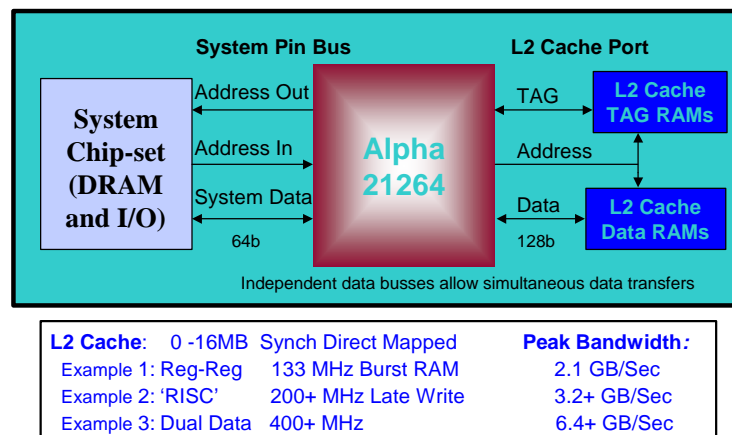
21264 Off-Chip Memory System Features

- **8 outstanding block fills + 8 victims**
- **Split L2 cache and system busses (back-side cache)**
- **High-speed (bandwidth) point-to-point channels**
clock-forwarding technology, low pin counts
- **L2 hit load latency (load issue to consumer issue) = 6 cycles + SRAM latency**
- **Max L2 cache bandwidth of 16 bytes per 1.5 cycles**
6.4 GB/sec with a 400Mhz transfer rate
- **L2 miss load latency can be 160ns (60 ns DRAM)**
- **Max system bandwidth of 8 bytes per 1.5 cycles**
3.2 GB/sec with a 400Mhz transfer rate

REK August 1998

23

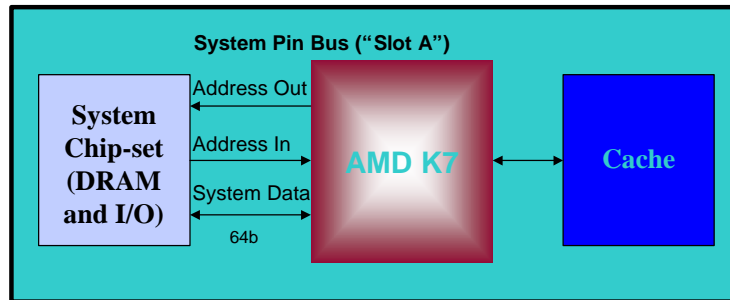
21264 Pin Bus



REK August 1998

24

Compaq Alpha / AMD Shared System Pin Bus (External Interface)

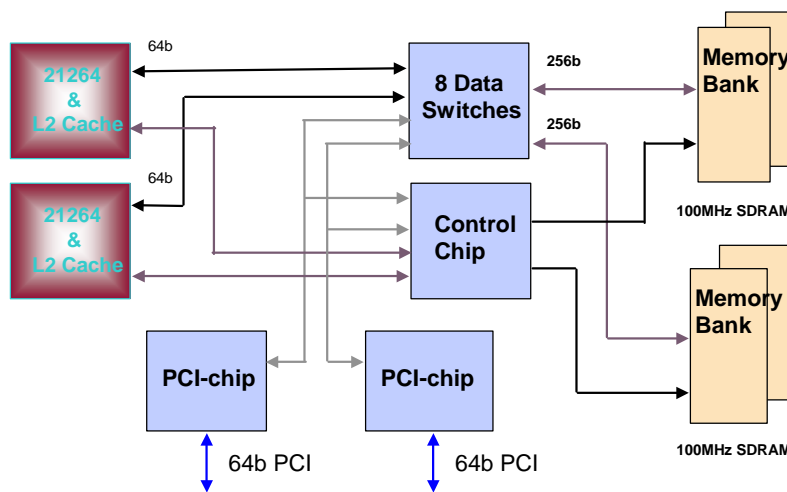


- High-performance system pin bus
- Shared system chipset designs
- **This is a win-win!**

REK August 1998

25

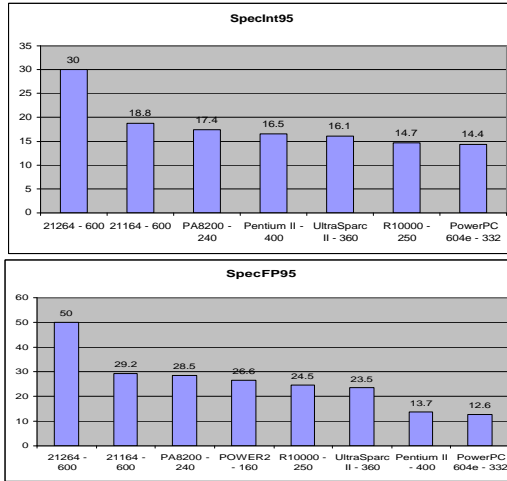
Dual 21264 System Example



REK August 1998

26

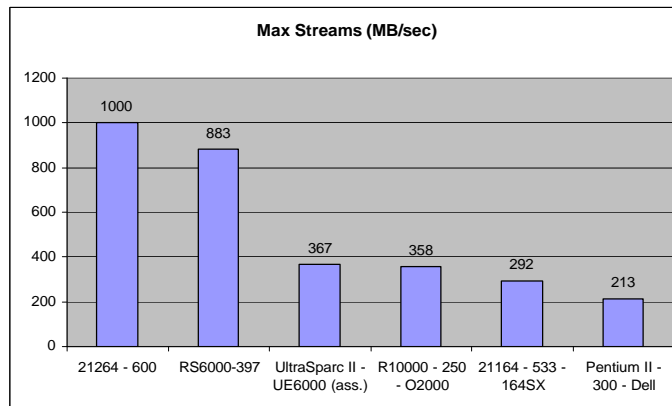
Measured Performance: SPEC95



REK August 1998

27

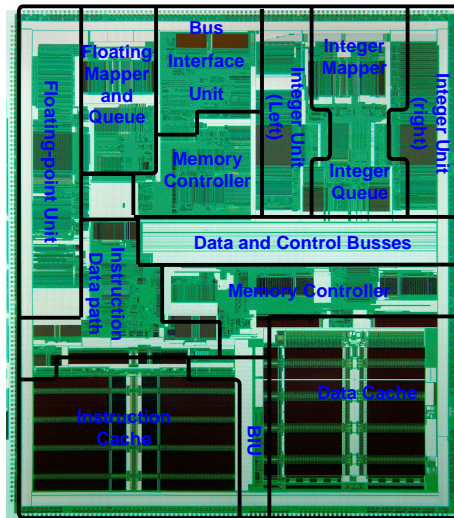
Measured Performance: STREAMS



REK August 1998

28

Alpha 21264



REK August 1998

29

Summary

- **The 21264 will maintain Alpha's performance lead**
 - 30+ Specint95 and 50+ Specfp95
 - 1+ GB/sec memory bandwidth
- **The 21264 proves that both high frequency and sophisticated architectural features can coexist**
 - high-bandwidth speculative instruction fetch
 - out-of-order execution
 - 6-way instruction issue
 - highly parallel out-of-order memory system

REK August 1998

30