MasPar MP-2 PE Chip: A Totally Cool Hot Chip

Won Kim, VLSI Architect and Designer
Russ Tuck, Systems Architect

MasPar Computer Corporation

749 N. Mary Ave., Sunnyvale, CA 94086
(408) 736-3300, FAX: (408) 736-9560
email: tuck@maspar.com

Outline

- Goal
- Strategy: Replication
- Design
- System Performance
- Why Custom? Why not COTS?
- Future Directions

Goal:

Affordable Solutions for Data Intensive Problems

- High **Delivered** Performance
- Large Problems
- Affordable: roughly $100k to $1M

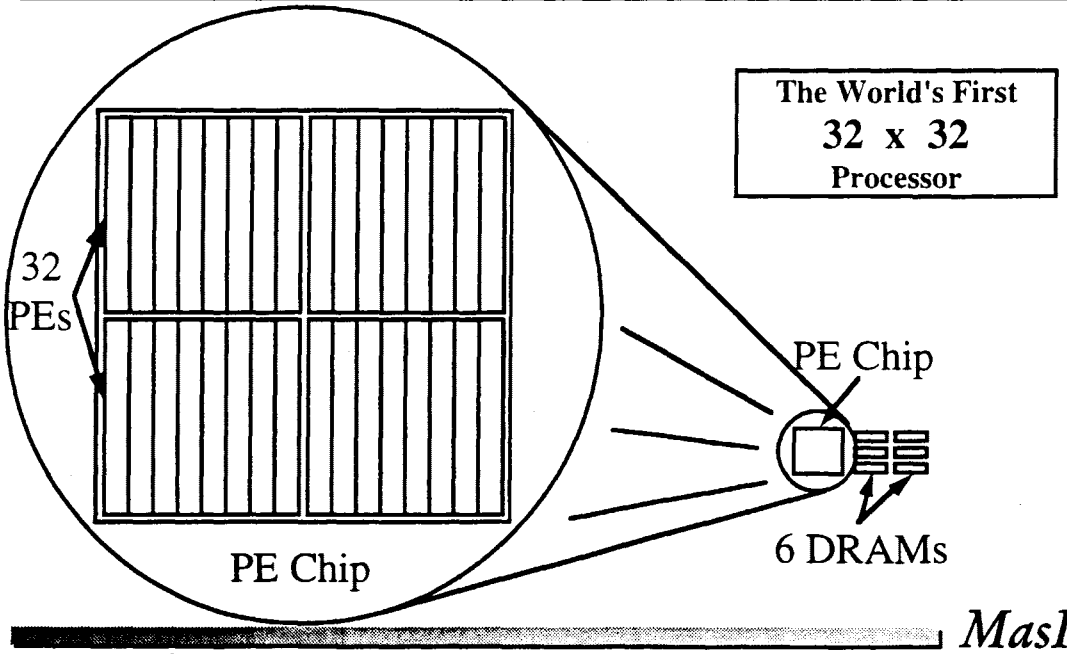
MasPar

Strategy: Replication

- Data Parallel
  - Programmable
  - Efficiently Use 1000s of Processors
- Synchronous
  - Reduces Communication Latency
  - Improves Communication Efficiency
  - Simplifies Communication Interfaces
- Balanced & Scalable
  - Communication Bandwidth & Latency
  - Memory Bandwidth

MasPar

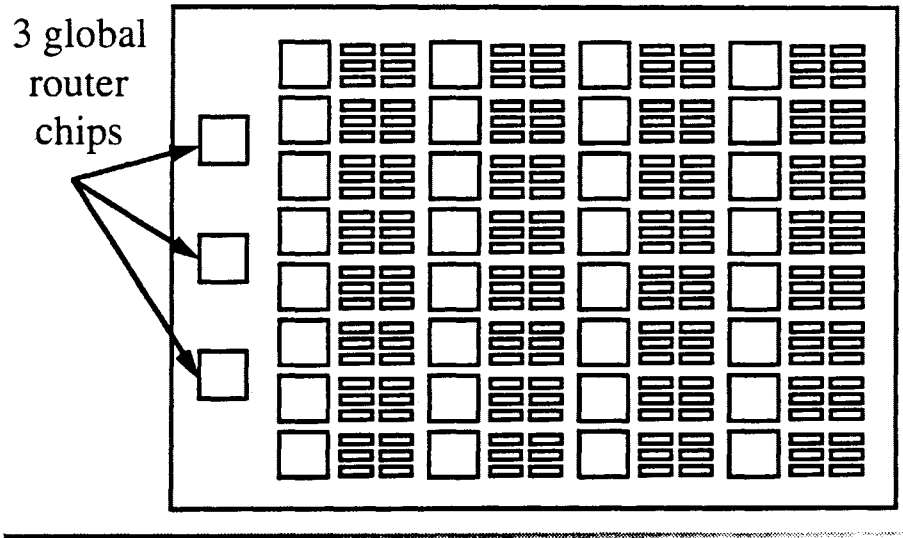
Replication: 32 PEs per Chip



MasPar

Replication: 32 Chips per Board

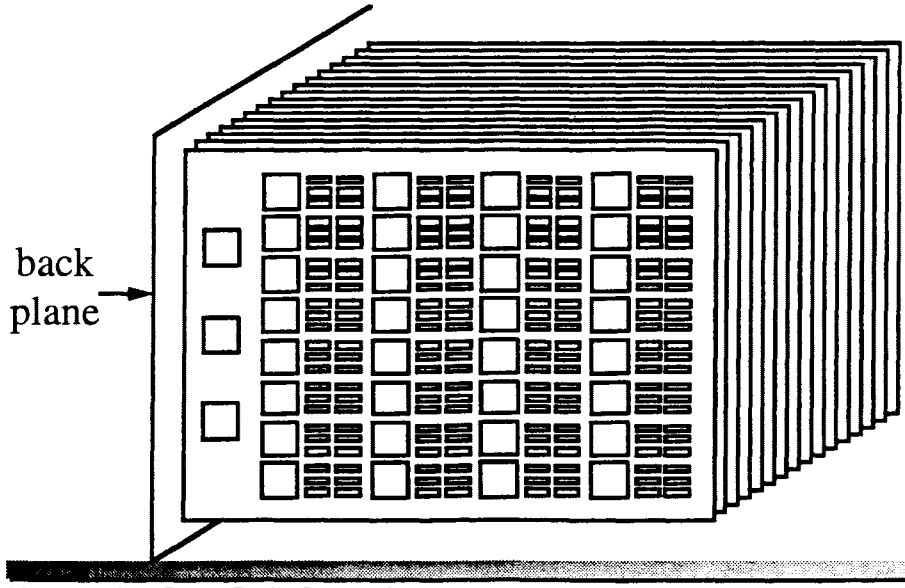
1024 PEs per Board



MasPar

Replication: 16 Boards per System

16K PEs per System



MasPar

Design

- Design leverage from replication
- Glueless interfaces
- Area and power efficient
- Optimized system performance

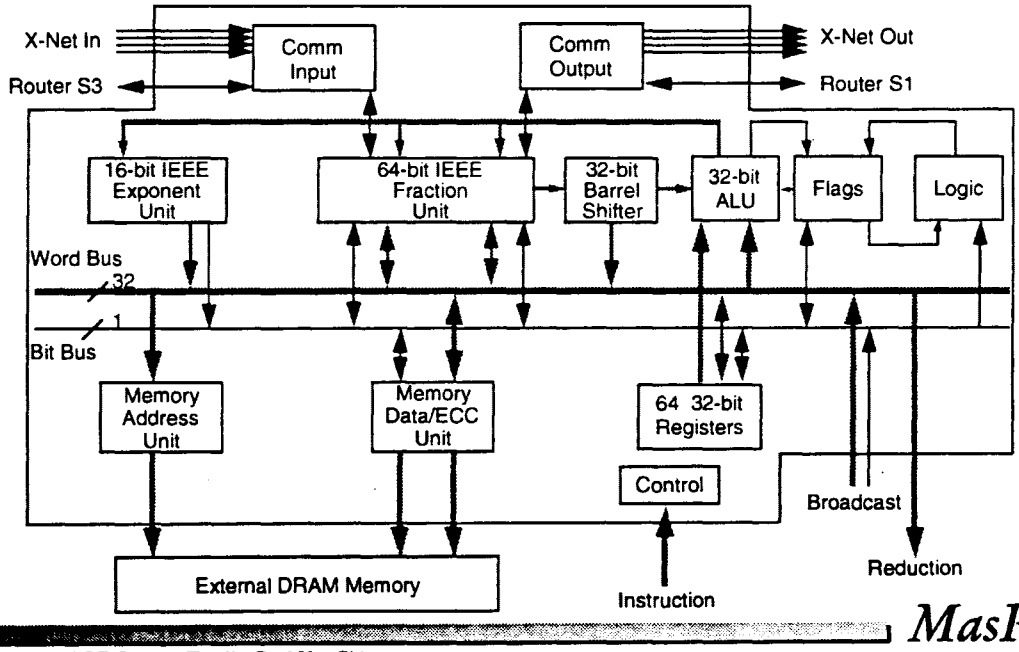
MasPar

PE "Slice" Design Leverage

- Design 1 PE
  - **Only 3% of chip area**
  - Optimize logic design
  - Hand-craft layout
- Replicate PE 32 times
- Rest of Chip
  - Shared Control
  - External Interface
  - About as many transistors as 1 PE

MasPar

PE Block Diagram



MasPar

PE: 32-bit Data Path

- 64 32-bit Registers
  - Special access modes support FP instructions
  - Some reserved for system use
- 32-bit ALU
- 32-bit Barrel Shifter
  - Built from power-of-2 shifters
    - Smaller and faster than "selection" shifters
    - Independently decodes shift amount in each PE

PE: 32-bit Data Path

- 64-bit IEEE Fraction Unit
  - Handles rounding and exception detection
  - Controls barrel shifter for renormalization
- 16-bit IEEE Exponent Unit
  - Controls barrel shifter for denormalization
- 1-bit Logic and Flag Units

PE: Interfaces

Memory

- Special Interface Registers
  - Address
  - Data
- Each PE can address its memory independently
- PEs can continue calculations during memory access
  - Array controller checks register dependencies automatically

PE: Interfaces

Communication

- Global Router
  - 16 PEs share 1 connection to external router
  - Inbound and outbound bit-serial ports
  - Arbitration and decoding logic independently select which PE uses inbound and outbound router port
  - Circuit Switched

PE: Interfaces

Communication (Cont.)

- Nearest-Neighbor
  - Inbound and outbound bit-serial ports
  - Choice of neighbor controlled globally
  - **Synchronous system**
    - low protocol overhead: 1 bit ("data valid")
- Broadcast
  - Copies 1 scalar value to all PEs
- Reduction
  - Global OR combines parallel data into scalar result

MasPar

PE Cluster Memory Interface

Memory bandwidth is key to delivering high performance on large problems.

- Efficient DRAM use
  - Can use every pin of every DRAM, all the time
  - Uses widest commodity DRAMS (x8)
  - Uses on-chip parallelism to automatically pipeline memory accesses
- Small, replicatable memory system
  - Each PE chip requires only 6 DRAMs + **no glue logic**
  - 16 PEs share 3 DRAMs that provide 16-bit data + ECC

MasPar

Circuit Design

Putting 32 PEs on one chip with a modest clock speed allows the use of circuit design methods that save area and power. This enables the replication strategy.

- Minimum device sizes everywhere except critical paths
- Flow-through logic computes a lot in one cycle
  - Avoids adding pipeline registers: saves power and area

Circuit Design (cont.)

- Decode and control done at chip edge, not in each PE
- Clocking: short phase 1 and long phase 2
  - time savings of 1-phase clock
  - simplicity of 2-phase clock
- Zero-detection in floating-point circuitry uses partial ripple carry with tree combining: saves space and power

Chip Statistics

- 950,000 Transistors
- 1.0µ two-level metal CMOS, 5.0 V
  - commodity process for affordable replication
- 64 Kbit SRAM register file
  - 20% of area
  - 50% of transistors
- 550mil x 550mil die size

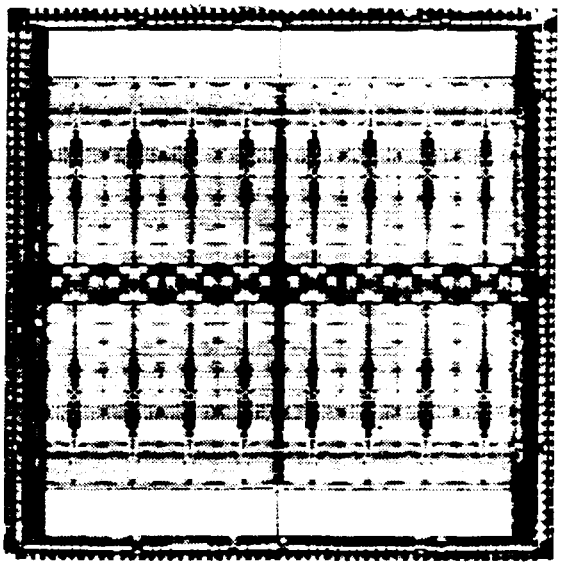
MasPar

Chip Statistics (cont.)

- 0.8 W -- low power:
  - simple air cooling
  - nexpensive plastic package
- 208 pin PQFP (plastic quad flat pack)
- 80 ns
  - modest clock rate saves power and area

MasPar

Die Photo



MasPar

Performance

Peak	Chip	System
Memory Bandwidth	45. MB/s*	23. GB/s
Nearest-Neighbor Bandwidth	45. MB/s	23. GB/s
Global Router Bandwidth	2.7 MB/s	1.3 GB/s
32-bit Integer	133. MOPS	68. GOPS
32-bit Floating Point	12.3 MFLOPS	6.3 GFLOPS
64-bit Floating Point	4.6 MFLOPS	2.4 GFLOPS

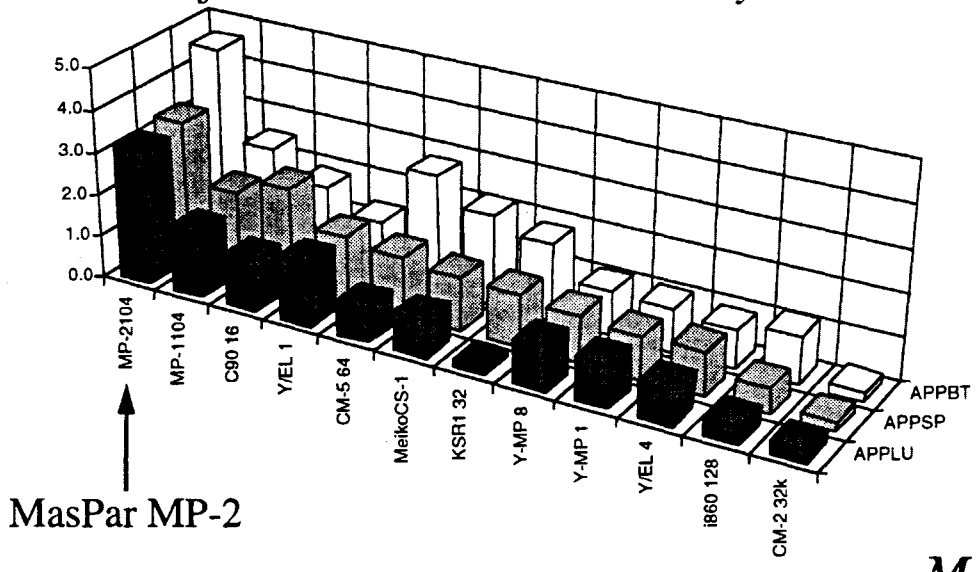
Linpack (64-bit) 1.6 GFLOPS

*Memory bandwidth achieved with only 6 DRAMs per PE chip

MasPar

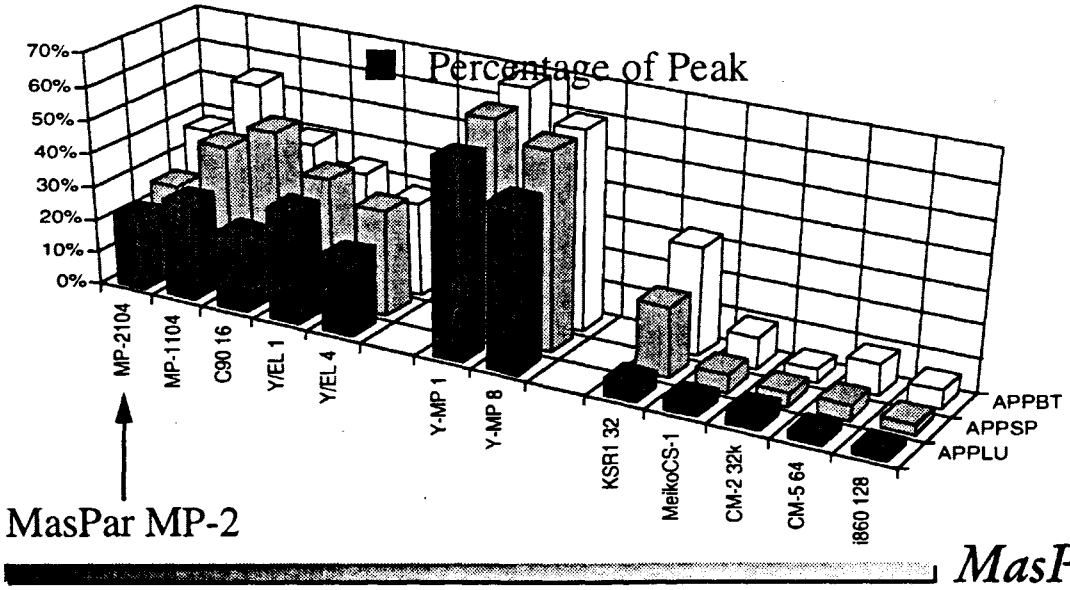
NAS Application Benchmarks

Performance/Price vs. 1 Cray Y/MP



MasPar

NAS Application Benchmarks



MasPar

Why Custom? Why not COTS*?

Philosophy:

- Leverage commodity parts when suitable
  - E.g., DRAM
- Use custom parts when give compelling advantage
  - E.g., PE Chip
    - communication latency & bandwidth
    - memory bandwidth
    - integration (glueless interfaces)
    - cost

* COTS = commodity off-the-shelf

MasPar

Future Directions

- More powerful processors
  - Borrow ideas from single-processor chips
- More processors
  - Already demonstrated scalability in hardware, architecture, and software

What other computer technology enjoys two such clear paths to continued performance growth?

MasPar