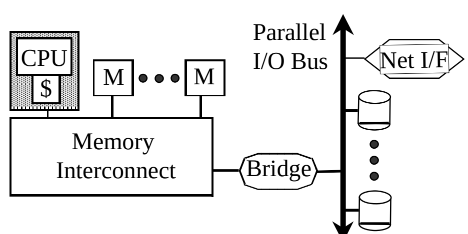
# An I/O Architecture for IRAM and SmartSIMM

Kimberly Keeton, Remzi Arpaci-Dusseau,
and David A. Patterson

UC Berkeley Computer Science Division
{kkeeton,remzi,patterson}@cs.berkeley.edu
http://iram.cs.berkeley.edu/

ISCA Workshop on Mixing Logic and DRAM
June 1, 1997

## Motivation: Limitations of Current System Architectures




Evolutionary Architecture

- ✶ I/O bus performance lags that of other system components
- ✶ Parallel I/O bus performance scaled by increasing clock speed and/or bus width
  - Ex: 32-bit PCI: ~50 pins; 64-bit PCI: ~90 pins
  - Implies even greater number of pins → greater packaging costs
- ✶ Are there alternatives to parallel I/O busses for IRAM?

2

## Motivation: I/O System for IRAM

- Communication advances: fast (Gbps) serial I/O lines [YangHorowitz96], [DallyPoulton96]
  - Serial lines require 1-2 pins per unidirectional link
  - Access to standardized I/O devices
    - Fibre Channel-Arbitrated Loop (FC-AL) disks
    - Gbps Ethernet networks
- Serial I/O lines a natural match for IRAM
- Benefits:
  - Avoids large number of pins for parallel I/O busses
  - IRAM can sink high I/O rate without interfering with computation
  - “System-on-a-chip” integration means chip can decide how to:
    - Notify processor of I/O events
    - Keep caches coherent
    - Update memory



## Motivation: Overview

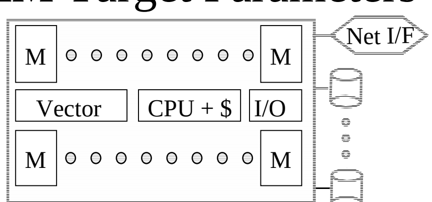
- How well will serial I/O work for IRAM?
  - Serial lines provide high I/O bandwidth for I/O-intensive app
  - Number of I/O lines is bottleneck for our chosen parameters
  - I/O bandwidth incrementally scalable by adding more lines
    - Number of pins required still lower than parallel bus
- How to overcome small memory capacity of single IRAM?
  - SmartSIMM: collection of IRAMs (and external DRAMs)
  - Can leverage high-bandwidth I/O to compensate for small memory
    - Demonstrate with external sorting case study
- In addition to other strengths, IRAM w/ serial lines provides high I/O bandwidth

## Outline

- ✶ Motivation
- ✶ IRAM and SmartSIMM
- ✶ Case Study:  External Sorting
- ✶ Future Work
- ✶ Conclusions



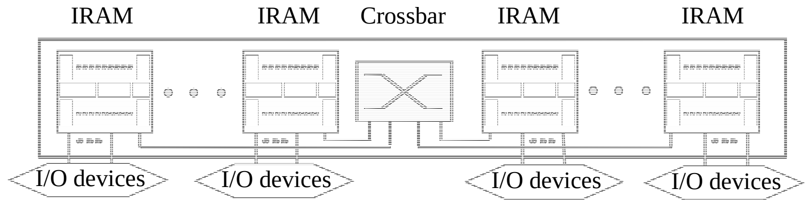
## Berkeley IRAM Target Parameters




| Characteristic | IRAM-I (1999) | IRAM-II (2002) |
|---|---|---|
| DRAM Generation | 256 Mbit | 1 Gbit |
| On-Chip Memory (MB) | 24 | 96 |
| On-Chip Memory B/W (GB/s) | 50 – 200 | 50 – 200 |
| I/O B/W via N Serial Lines (GB/s) | 0.5 – 2.0 | 0.5 – 4.0 |
| Individual Serial Line B/W (GB/s) | 0.25 | 0.5 |
| On-Chip Memory Latency (ns) | 20 – 30 | 20 – 30 |
| Processor Speed (MHz) | 300 – 500 | 500 – 1000 |
| Vector Performance (GFLOPS) | 4 | 16 |

+ High on-chip memory B/W, low on-chip memory latency
- Small on-chip memory capacity

# SmartSIMM




- ✶ Up to 16 IRAMS connected via high-speed crossbar
- ✶ May also be configured with mix of IRAM and fast DRAM (e.g., RAMBUS, synch. DRAM)
- ✶ Programmed as cluster or shared address space NUMA

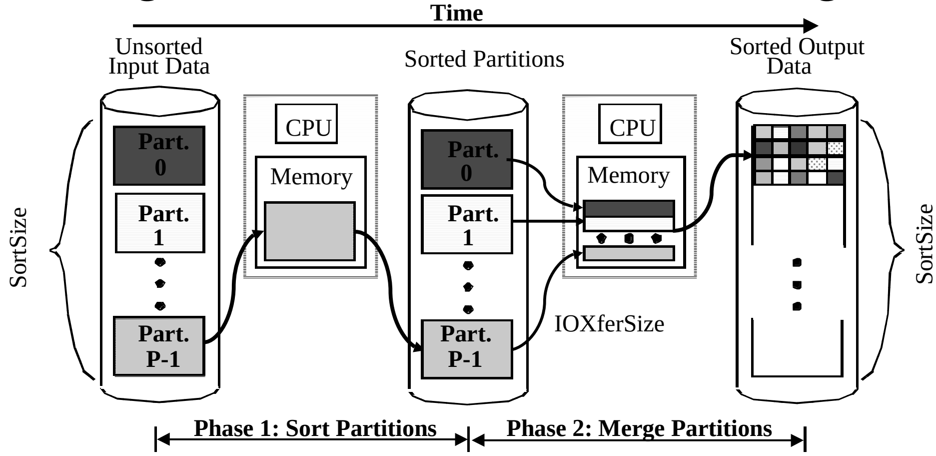
7

# Case Study: External Sorting

- ✶ External: data must start and end on disk
- ✶ Stresses I/O and memory systems
- ✶ Important to numerous research communities
  - Database, scientific
- ✶ MinuteSort: how much can we sort in a minute?
  - Benchmark designed by database community [Nyberg94]
  - Current record: 8.4 GB on 95 UltraSPARC I's w/ Myrinet [NOWSort:Arpaci-Dusseau97]
- ✶ Terminology:
  - One-pass sort: data fits into memory
  - Two-pass sort: data larger than memory
- ✶ Use two-pass sort to match IRAM's memory & I/O systems

8

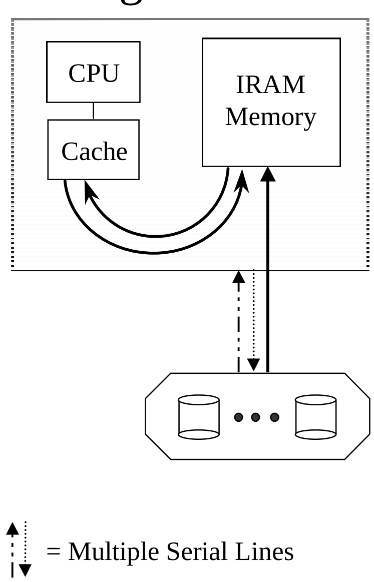
## Single Node Two-Pass External Sorting




- Size($P_i$) = MemorySize
- # Partitions (P) = SortSize/MemorySize
- Required disk I/O: D = 4*SortSize GB / (1 min)

9

## Single Node Sort: Analysis




- Read Phase Steps
  - Read from disk
  - Examine keys
  - Scatter keys into sort buckets
- Bandwidth Analysis [NOWSort:Arpaci-Dusseau97]
  - Required disk B/W = D GB/s
  - Required memory B/W = 3*D GB/s
- SortSize
  - SortSize=MemorySize$^2$/IOXferSize
- Analytical methodology:
  - Combine these requirements with IRAM technological constraints
  - Assume computation is free

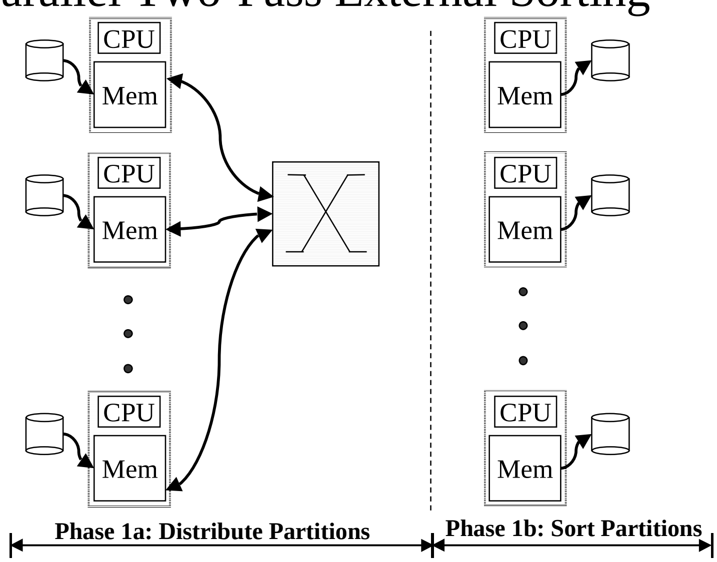
10

## Single Node Results: IRAM-I (1999)

- Total MinuteSort amount: 9.0 GB
- IRAM-I configuration:
  - Total on-chip memory: 24 MB
  - Disk bandwidth: 0.6 GB/s
  - Number of disks: 46
  - Number of 2 Gbps serial I/O lines: 6 (bottleneck)
  - Memory bandwidth: 1.8 GB/s
- Comparison point: IRAM-I with 64-bit, 66 MHz PCI
  - MinuteSort amount: 2.2 GB
    - ~4X lower than serial line analog
  - More pins: ~80 for PCI vs. 16 - 32 for serial lines
- 1997 reference sort amount: 8.4 GB by 95 UltraSPARCs

11

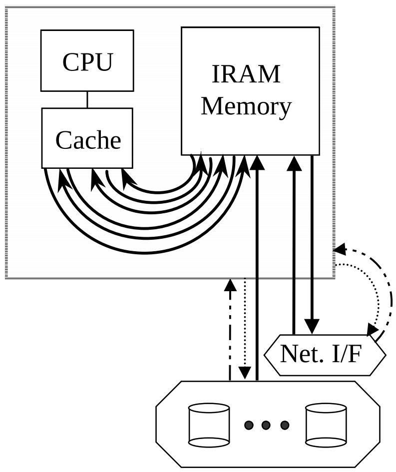
## Parallel Two-Pass External Sorting




- Phase 2 same as single-node case (completely local)

12

# Parallel Sort: Analysis




= Multiple Serial Lines

- Read and Communicate Steps
  - Read from disk
  - Examine records
  - Scatter records to send buffers
  - Copy keys to send buffers
  - DMA keys out
  - Receive keys
  - Examine keys
  - Scatter keys into sort buckets
- Bandwidth Analysis
  - Required disk B/W = D GB/s
  - Required network B/W = 2*D GB/s
  - Required memory B/W = 6 to 9*D GB/s
- SortSize: N*(per-node SortSize)

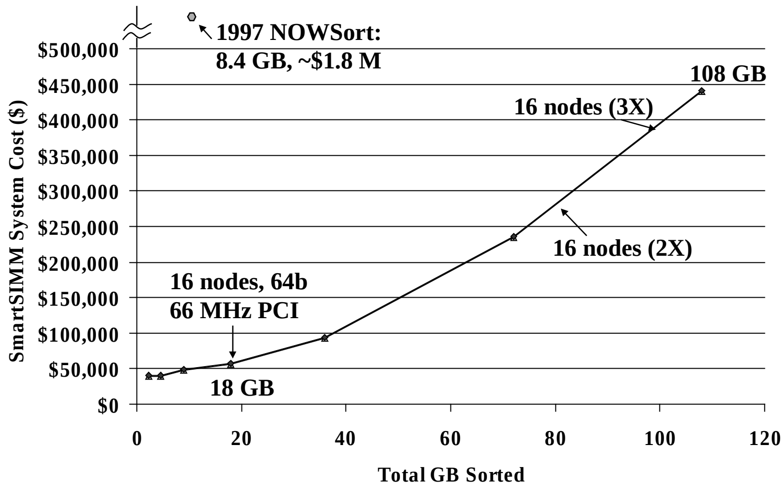
13

# Estimated Costs for SmartSIMMs

| **Quantity** | **1997** | **1999 IRAM-I (256 Mbit DRAM)** | **2002 IRAM-II (1 Gbit DRAM)** |
|---|---|---|---|
| Memory Cost / MB (from SIMM) | $5 | $3 | $1 |
| DRAM Chip Capacity (MB) | 8 | 32 | 256 |
| Est. Min. IRAM Cost (2X DRAM) | N/A | $185 | $325 |
| Est. Max. IRAM Cost (3X DRAM) | N/A | $275 | 485 |
| SmartSIMM Crossbar Cost | N/A | $80 | $60 |
| Disk Cost / MB | $0.27 | $0.10 | $0.02 |
| Disk Capacity (GB) | 2.2 | 5.5 | 22.5 |
| 8-Disk Enclosure | $1900 | $1700 | $1500 |

14

# Parallel Results:  SmartSIMM-I (1999)




15

# SmartSIMM-I Parallel Results Summary

- SmartSIMM-I MinuteSort amount:  108 GB
- Required per-node bandwidths:
  - Disk:  0.45 GB/s (24 disks)
  - Network:  0.9 GB/s
  - Memory:  5.4 GB/s
- Limitation:  number of serial lines
- SmartSIMM-I cost:  $440K - $441K
  - System cost  relatively insensitive to IRAM cost
  - Disk subsystem is 99% of overall cost
- Comparison point:  IRAM-I with 64-bit, 66 MHz PCI
  - MinuteSort amount: 18 G
    - ~5X lower than serial line analog
  - More pins:  ~80 for PCI  vs. 16 - 32 for serial lines

16

## Future Work

- Vectorize sort application
  - Vectorized sort algorithms known [ZaghaBlelloch91]
- Other candidate applications
  - Decision support database applications
  - Database transaction processing
  - Image and video transcoding
  - System-level applications
    - TCP/IP checksumming
    - RAID parity calculations
    - Data copying
    - Data movement
    - Memory zeroing

17

## Conclusions

- IRAM with high-speed serial lines provides high B/W I/O
- SmartSIMM: collection of IRAMs (and external DRAM)
- External sorting lessons
  - Can trade off high I/O bandwidth for low IRAM memory capacity
  - Single-node IRAM-I MinuteSort
    - Number of I/O serial lines is limiting factor
  - 16-IRAM SmartSIMM-I MinuteSort
    - Number of I/O serial lines is limiting factor
  - Can easily scale I/O performance by adding serial lines
    - Still keeps pin count lower than parallel alternative
  - SmartSIMM provides cost-effective sorting solution
    - Cost of system is essentially cost of disks

18

## Backup Slides

19

## Berkeley IRAM Target Parameters

| **Characteristic** | **IRAM-I (1999)** | **IRAM-II (2002)** |
|---|---|---|
| DRAM Generation | 256 Mbit | 1 Gbit |
| On-Chip Memory (MB) | 24 | 96 |
| On-Chip Memory B/W (GB/s) | 50 – 200 | 50 – 200 |
| I/O B/W via N Serial Lines (GB/s) | 0.5 – 2.0 | 0.5 – 4.0 |
| Individual Serial Line B/W (GB/s) | 0.25 | 0.5 |
| On-Chip Memory Latency (ns) | 20 – 30 | 20 – 30 |
| Processor Speed (MHz) | 300 – 500 | 500 – 1000 |
| Vector Performance (GFLOPS) | 4 | 16 |
| Power Consumption (W) | 0.5 – 2.0 | 0.5 – 2.0 |
| Die Size ($mm^2$) | 300 | 420 |

20

## Technology Trend Assumptions

| | **Disk Capacity (GB)** | **Media Transfer Rate (MB/s)** | **Track Buffer Xfer Rate (MB/s)** | **Access Time (ms)** | **Serial Line Speed (Gbps)** |
|---|---|---|---|---|---|
| Annual Relationship | 1.6 | 1.4 | 1.4? | 0.7 | ? |
| 1997 [Seagate97] | 2.2 | 15 | 100 | 5.07 | 1 |
| 1999 (est.) | 5.5 | 29 | 196 | 2.48 | 2 |
| 2002 (est.) | 22.5 | 80 | 540 | 0.85 | 4 |

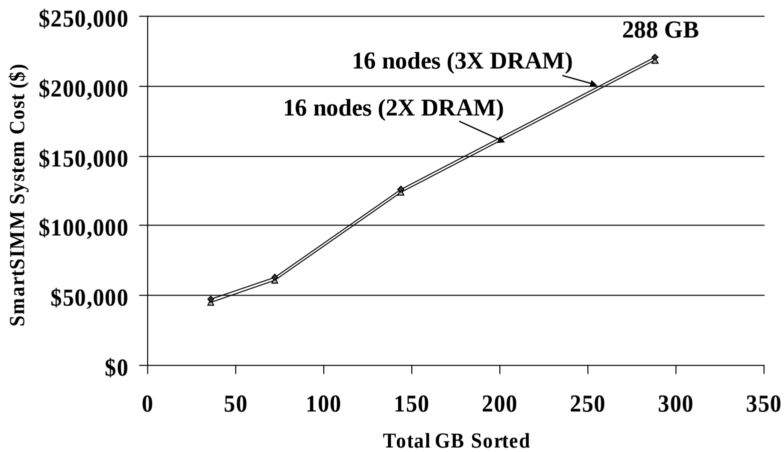
21

## Single Node Results: IRAM-II (2002)

- Total MinuteSort amount: 36.6 GB
- IRAM-I configuration:
  - Total on-chip memory: 96 MB
  - Disk bandwidth: 2.4 GB/s
  - Number of disks: 48
  - Number of 4 Gbps serial I/O lines: 8 (bottleneck)
  - Memory bandwidth: 7.3 GB/s
- 1997 reference sort amount: 8.4 GB by 95 UltraSPARCs

22

# Parallel Results: SmartSIMM-II (2002)




✶ Nearly threefold increase in MinuteSort amount over SmartSIMM-I
✶ SmartSIMM cost relatively insensitive to IRAM cost
✶ IRAM cost: 2% to 4% of overall system cost

23

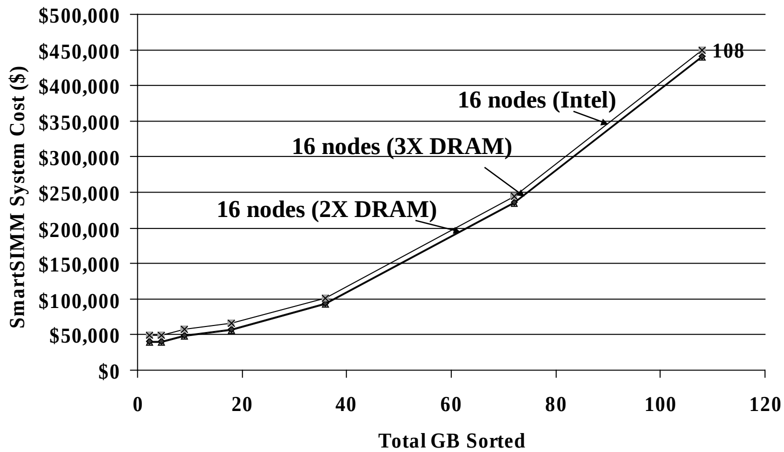
# SmartSIMM-II Parallel Summary

✶ SmartSIMM-II MinuteSort amount: 288 GB
✶ Required per-node bandwidths:
- Disk: 1.2 GB/s (18 disks)
- Network: 2.4 GB/s
- Memory: 10.8 GB/s

✶ Limitation: number of serial lines
✶ SmartSIMM-I cost: $218K - $221K
- System cost relatively insensitive to IRAM cost
- Disk subsystem is 96 - 98% of overall cost

24

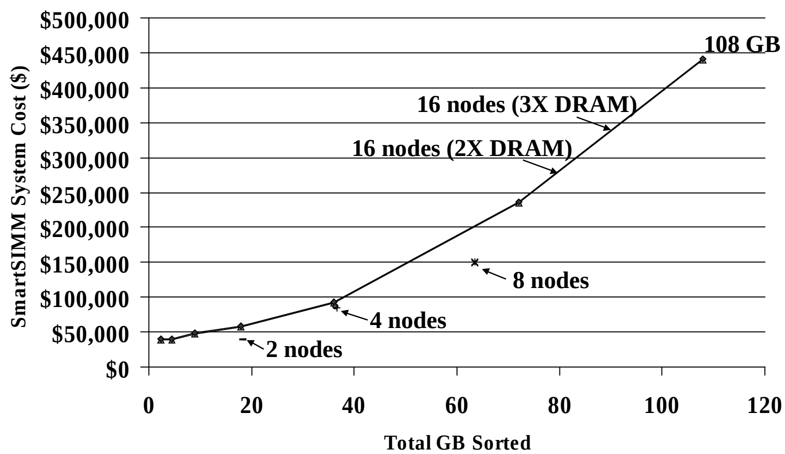
## SmartSIMM-I Sensitivity to IRAM Cost




- ✶ Intel: assume IRAM costs as much as latest Intel processor (reasonable??)
- ✶ SmartSIMM system cost relatively insensitive to IRAM cost
- ✶ IRAM cost: 1% to 3% of overall system cost

25

## SmartSIMM-I Sensitivity to Configuration




- ✶ Compare 16-IRAM costs with smaller node configurations
  - IRAM-I used as external DRAM in smaller node configurations
- ✶ Smaller configurations don’t achieve peak performance of 16-node case
- ✶ For their sort amounts, smaller configs more cost effective

26

## One-Pass Sort for 16-Node PC Cluster

- MinuteSort amount: 64.0 GB
- Required per-node bandwidths:
  - Disk bandwidth: 0.13 GB/s (5 disks)
    - D = 2*SortSize / (1 min)
  - Network bandwidth: 0.27 GB/s
  - Memory bandwidth: 1.20 GB/s
- Limitation: memory capacity
- System cost: $517K
- Cost breakdown:
  - Processor and memory: 82%
  - Disk subsystem: 14%
  - Network infrastructure: 4%

27

## Two-Pass Sort for 16-Node PC Cluster

- MinuteSort amount: 38.4 GB
- Required per-node bandwidths:
  - Disk bandwidth: 0.16 GB/s (6 disks)
    - D = 4*SortSize / (1 min)
  - Network bandwidth: 0.32 GB/s
  - Memory bandwidth: 1.44 GB/s
- Limitation: PCI bandwidth
- System cost: $526K
- Cost breakdown:
  - Processor and memory: 80%
  - Disk subsystem: 15%
  - Network infrastructure: 5%

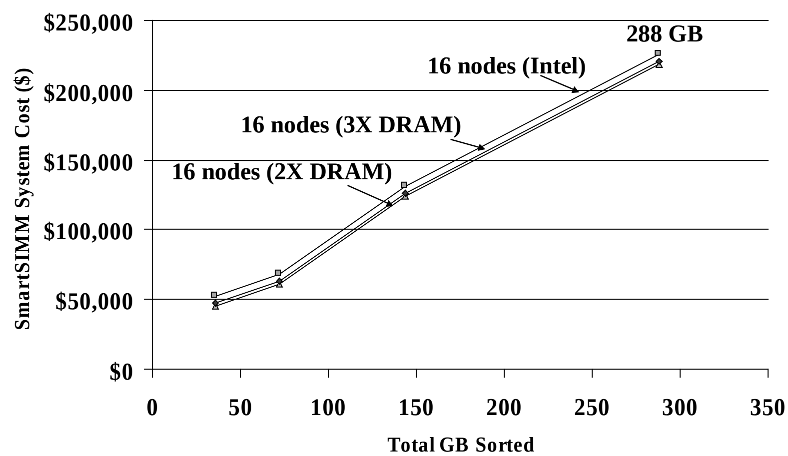
28

# PC Technology and Cost Parameters

- ✶ Main memory: 4 GB, from 256 Mbit DRAM
- ✶ Parallel I/O bus: 64-bit, 66 MHz PCI
- ✶ Disk assumptions: same as SmartSIMM-I study
- ✶ Other estimated costs:
  - 16-port Myrinet switch: $1800
  - Myrinet network interface: $1200
  - Myrinet high-speed cable: $125
  - Intel-like processor: $775

29

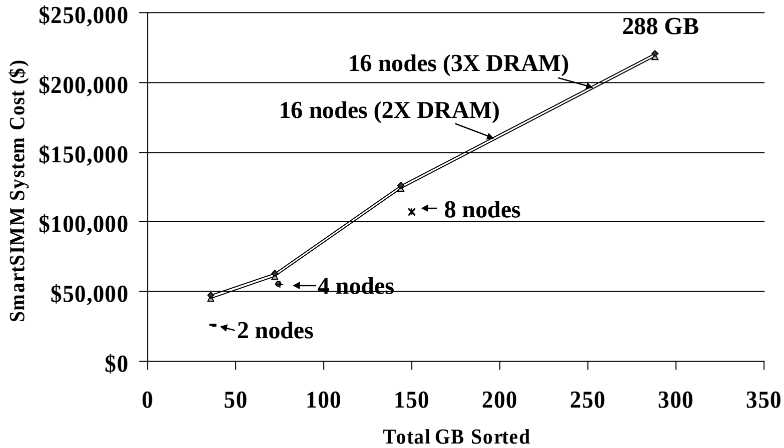
# SmartSIMM-II Sensitivity to IRAM Cost




- ✶ Intel: assume IRAM costs as much as latest Intel processor (reasonable??)
- ✶ SmartSIMM cost relatively insensitive to IRAM cost
- ✶ IRAM cost: 2% to 6% of overall system cost

30

## SmartSIMM-II Sensitivity to Configuration




- ✶ Compare 16-IRAM costs with smaller node configurations
  - IRAM-II used as external DRAM in smaller node configurations
- ✶ Smaller configurations don't achieve peak performance of 16-node case
- ✶ For their sort amounts, smaller configs more cost effective

31

## Derivation of SortSize

- ✶ Phase 1 constraint:
  - P = number of partitions = SortSize / MemorySize
- ✶ Phase 2 constraint:
  - MemorySize = P*IOXferSize
- ✶ Solving simultaneously for SortSize:
  - SortSize=MemorySize$^2$/IOXferSize

32

## Serial Lines for Conventional Systems?

- Serial lines in use for connecting multimedia I/O peripherals (e.g., Firewire, IEEE 1394)
- Small pin count argument not as compelling
  - Multi-chip conventional systems already designed with many parallel interfaces
- Serial line can't plug directly into memory crossbar
  - Serial line = point-to-point link
  - Must be able to:
    - DMA data to memory
    - Maintain cache coherency
    - Notify processor of I/O events
  - Need to build I/O device controller into memory crossbar?
- Potential architecture:
  - Separate serial line controller that plugs into memory crossbar port
  - Unlikely in forseeable future?

33

## Serial Line Power Consumption

- Serial line power consumption [Seitz97], [Jeong97]:
  - 150 mW - 200 mW per Gbps for 0.5 to 0.35 $\mu$m CMOS
  - Target: 100 mW per Gbps
- Possible to conditionally operate serial lines?
  - Ex: network lines in 2-pass sort
    - Majority of lines for parallel 2-pass sort
    - Not used for (local) 2nd phase

34

## Serial Line Area Requirements

* Approx. area parameters for 0.35 μm CMOS [Jeong97]:
  * Central, shared PLL: 0.5mm x 0.5mm
  * Transmitter: 0.2mm x 0.5mm
  * Receiver: 0.5mm x 1.0mm
* Scaling to 0.25 μm CMOS:
  * 50 - 60% area-wise
  * 70% length-wise

35