SF 1008-RA (FRONT 7-2003)
Supersedes (3-2003) issue

Information Release

SAND 2003-2869 P

REVIEW & APPROVAL FORM

Originating organization: Please complete Sections 1 - 6. Print or type all information. See attached instruction sheets for additional information.

This form is used to review and approve information releases before they are released outside of Sandia.

Public releases must go through the Formal R&A Process, in which case this form must be completed through Section 10.
For releases going through the Organizational R&A Process, organizational management is encouraged to complete this form through Section 6 and to file it for future reference.
This form can also be used (through Section 6 & Section 8) to document approvals when an information release changes distribution limitations (e.g., when Internal Distribution Only becomes Unlimited Release).
For information on which R&A process to use, or additional R&A information:

- See the "Review and Approval for Communication" webpage: http://www-irn.sandia.gov/recordsmgmt/revapprov/revapprov.htm or
- Contact Linda Cusimano (NM), (505) 844-4980 ; Kelly McClelland (CA), (925) 294-2311

SECTION 1. Protecting Sandia and Partnership Interests.

Is this release the result of □ a CRADA? □ Work for Others? □ Other partnership, MOU agreement, funding source, or understanding OF ANY KIND? □ Information controlled by other agencies

☒ If NO: Go to Section 2.
□ If YES: Agreement Number is

Has your partner, non-SNL information owner, or funding agency given approval for this release? □ Yes □ No

SECTION 2. Document Title and Author Information:

Full title of document **How Sandia May Reach 1000 TFLOPS**

Author or contact (Sandian) **Erik P. DeBenedictis** Signature
(Full First Name Full Middle Name Full Last Name)

Phone No. **505 284 4017** E-Mail Address **epdeben@sandia.gov** Org. No. **9223** Mail Stop No. **1110**

Project number **7101** Task number **13.04** (Identifies funding source – will not be charged)

□ Contract Author to Sandia. (Contractor's name and contract no.)

SECTION 3. Document Format and Release Event Information. Indicate the planned format(s) of the information release, as well as information about the release event.

Document Format(s): □ SAND Report □ Abstract □ Conference Paper □ Journal Article □ Sandia Open Network (External) □ Computer Software ☒ Publication (all other types of publications including reports, vugraphs, posters, exhibits, displays, videos, brochures, internal memoranda, newsletters, factsheets)

Release Event: Indicate the name of the conference, meeting, or publication, the sponsoring organization, and the place and date of event. If this release is an electronic posting, provide the current viewing address and intended posting location.

Name of Conference / Journal / Book: **Mission Critical Computing Conference**
Place of Event: **Washington DC** Date: **6/19/03 thru 6/20/03**
Internet Address of Electronic Posting:

SECTION 4. Classification and Sensitivity of Information. Contact Classification Dept. 3132 (8511 - CA) for questions.
Indicate classification level and category of information release or whether information release is unclassified:
Classification of: Document Title U Document Abstract NA The Document U

□ Classified - Limited Release. Indicate additional access restrictions:
□ NWD Sigma ______ □ CNWDI □ NOFORN □ Program Designated Special Handling (Distribution Limitation) □ Other ______

□ Unclassified - Limited Release. Indicate all Unclassified Controlled Information (UCI) categories access restrictions:
□ Export Controlled Information (ECI) ITAR/EAR/______
□ Non-Sandia Proprietary Information
□ Official Use Only (OUO) Exemption No. ______
□ Patent Caution
□ Protected CRADA Information (Release date ______)
□ Program Designated Special Handling (Distribution Limitation)
□ Unclassified Controlled Nuclear Information (UCNI)
□ Other (specify) ______

☒ Unclassified - Unlimited Release. Information is unclassified with no access restrictions, i.e., distribution may be made worldwide.

□ **DUSA Exemption -** This information is released under DUSA Exemption ______ (Mark appropriate sensitivity above. **Section 8 review not required.**)

Derivative Classifier (DC) who is knowledgeable of information sensitivity or DUSA Delegate: PAUL YARRINGTON Paul Yarrington 9230 7/30/03
Name Signature Org. Date

(Continued on page 2)

2003-2869P

SF 1008-RA (7-2003)
Supersedes (3-2003) issue

SECTION 5. Disclosure of Technical Advance

A Technical Advance is an original achievement or non-obvious progress in a scientific or engineering sense, including the creation of software. It may be protected by patent or copyright. The Originators of a Technical Advance may be inventors or authors.

Does the subject of this Information Release represent a Technical Advance as defined above?
□ Yes ☒ No If **No**, go to Section 6.

If **Yes**, has a Disclosure of **Technical Advance (TA)**, Form SF 1155-TD, been filed with the Sandia Patent and Licensing Center?
□ Yes SD No. ______________________ □ No If No, please follow up with a TA form obtainable from:

SECTION 6. Line/Program Signatures and Approvals.

Print or type all author information; obtain appropriate signatures from next-level manager. Where concurrence is obtained in case of multiple authors, approval need only go through the principal author's line organization.

Authors' Names (Print or type) (Full First, Middle, & Last Name)	Org. No./ Mail Stop	Phone No.	Next Level Manager's Signature	Date
Erik P. DeBenedictis (Full First Name Full Middle Name Full Last Name)	1110	505 284 4017	Neil Pundit	7/24/03
(Full First Name Full Middle Name Full Last Name)				
(Full First Name Full Middle Name Full Last Name)				
(Full First Name Full Middle Name Full Last Name)				
(Full First Name Full Middle Name Full Last Name)				

Program Manager's Name and Signature ______________________

THE FOLLOWING SECTIONS ARE USED FOR THE FORMAL REVIEW & APPROVAL PROCESS.

SECTION 7. To be completed by the Patent and Licensing Department (NM: 11500/MS 0161, CA: 11600/MS 9031).*

Copyright Interest? □ Yes ☒ No If Yes, copyright may be asserted, subject to DOE approval.

Patent Interest? □ Yes ☒ No If Yes, TA form has been or should be submitted.

Patent Caution? □ Yes ☒ No If Yes, TA form has been or should be submitted and dissemination will be limited.

Patent Attorney's/Agent's Signature ______________ Date 7/31/03

*** Only required for Technical and Scientific Information releases.**

SECTION 8. To be completed by the Classification and Sensitive Information Department (NM: 3132/MS 0175, CA 8511, MS 9021).

Signature ______________ Date 4 Aug 03

SECTION 9. To be completed by the Promotional Communications Departments (NM: 9612/MS 0612, CA: 8815/MS 9021).

Promotional Communications must be reviewed for adherence to Corporate "Common Look and Feel" guidelines by PR & Communications Center 12600, MS 0619 (NM) or by Communications & Public Affairs Dept. 8528, MS 9131 (CA). In addition, all publications (including SAND Reports and fact sheets) distributed outside of Sandia that use color (ink) **printing**, as well as all internal products of two or more colors that use color **printing**, must go through a Section 9 review. **NOTE:** SAND Reports and internal release products that use color **copying** do **not** need a Section 9 review. (Contact Printing & Duplicating Dept. 12630 to get a head start on DOE approval.)

DOE approval received __________ Date Signature ______________ Date __________

SECTION 10. To be completed by the Review & Approval Desk (NM: 9612/MS 0612, CA: 8511/MS 9021).

Approved under the following conditions:

1. **That the following statement is printed on the document: Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.**
2. **That, if the release has been determined to be sensitive or classified, all the appropriate markings are on the released item;**
3. **That all edits requested by reviewers are completed before release;**
4. **That the final version is submitted to the Technical Library.**

Signature Dorothy Martin Date AUG 5 2003

How Sandia May Reach 1000 TFLOPS

Erik P. DeBenedictis

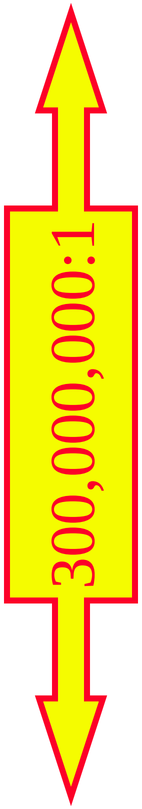
SAND 2003-2869P

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy under contract DE-AC04-94AL85000.

How To Reach 1000 TFLOPS?

- Time Acceleration
  - Scaling is mathematical based on # CPUs
  - Time is physical based on clock rate
  - If you keep architecture and # CPUs the same but increase the clock rate, the speed goes up, and efficiency stays the same
    - Scaling has to be right

- Has It Been Done Before?
  - Cosmic Cube, 1981
    - 64×50 KFLOPS
  - nCUBE 10, 1988
    - 1024 × 1 MFLOPS
  - nCUBE 2, 1990
  - Paragon, 1995
  - ASCI Red, 1998-
    - 9960 × 230 GFLOPS
  - ASCI Red Storm, 2004-
    - 10368 × 4 GFLOPS
  - Petaflop?



300,000,000:1

How to Spec the Machine?

- If the Government sector specifies the machine
  - It will be a linear speedup over ASCI Red
  - We will be able to predict performance
  - Project management will cost a bundle

- If industry designs the machine
  - It will have creative improvements designed to improve commercial potential
  - Untested improvements introduce risk
  - We are unlikely to be able to predict performance

Red Storm Scaling to 1000 TFLOPS

- Peak FLOPS 40 T → 1000 T (25x)
- Per node 4 GFLOPS → 100 GFLOPS
- Memory capacity stays about same at 1 byte/FLOPS
- Memory bandwidth 4 bytes/FLOPS stays the same
- Network bandwidth 4 bytes/FLOPS stays the same
- Latency (local/global) 2 µs/5 µs → 80 ns/200 ns

Ug

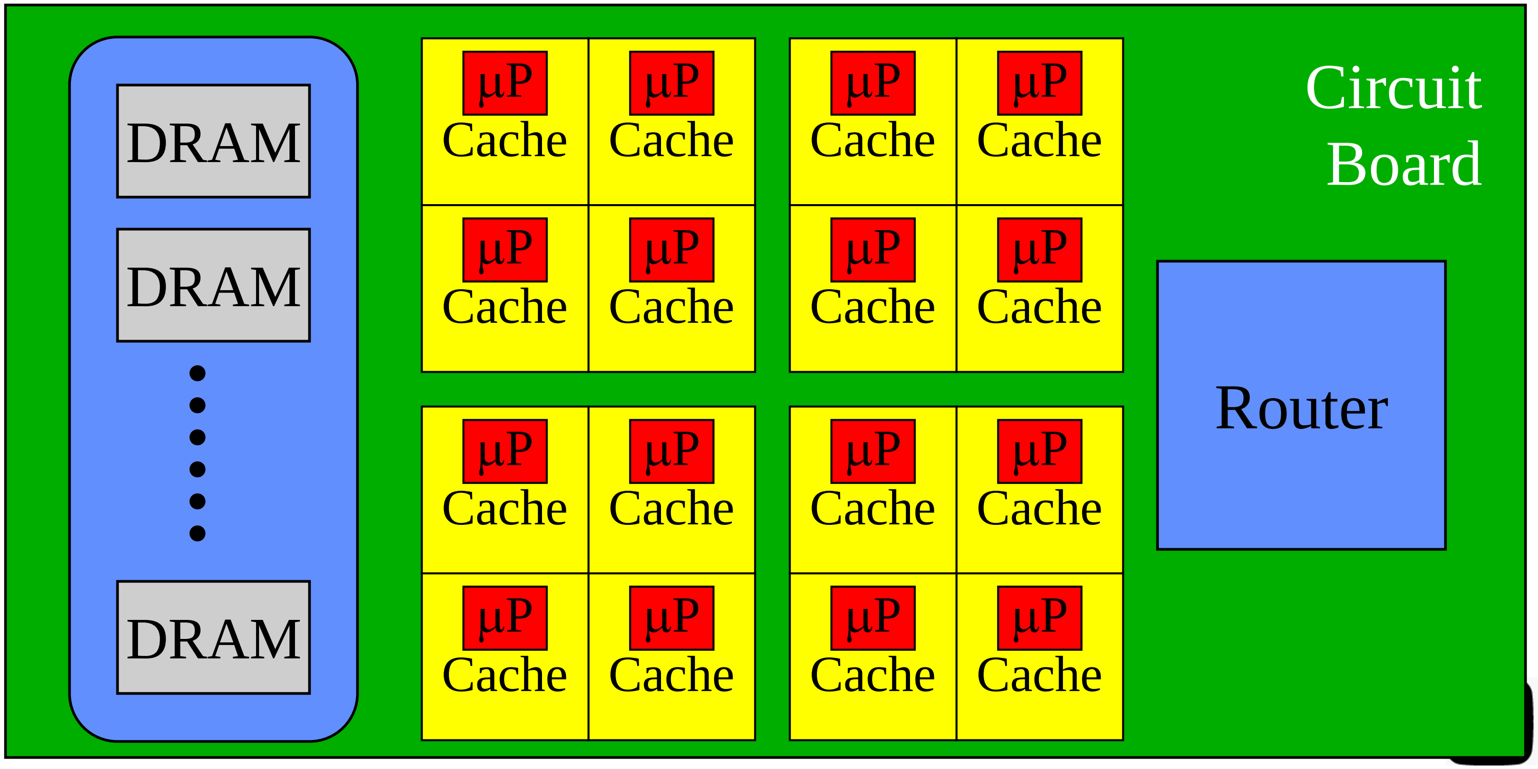
Much more Later

- Risk Factor
  - 100 GFLOPS CPU must be a SMP because cores cannot run this fast
  - However, various SMP nodes work OK up to n=8-16 (ASCI White, etc.)

Balance: time to operate on a number ~ time to send number across machine

Processors

- **Fred Weber (AMD) @ Salishan**
  - **"2008 144 GFLOP 4P * 4 * 9 GHz"**



DRAM

DRAM

DRAM

µP Cache | µP Cache | µP Cache | µP Cache

µP Cache | µP Cache | µP Cache | µP Cache

µP Cache | µP Cache | µP Cache | µP Cache

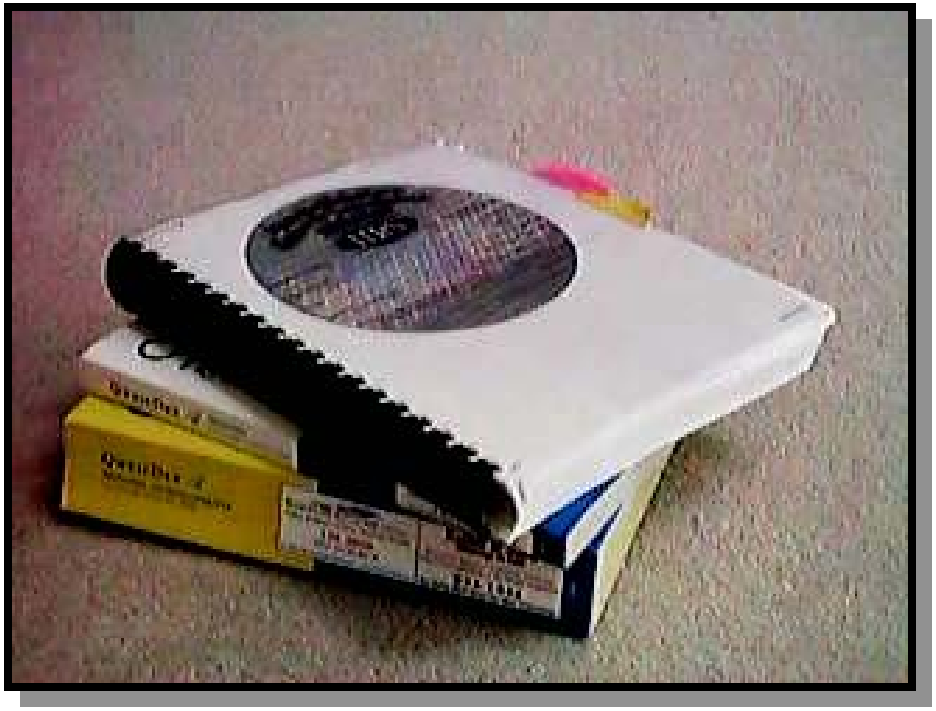
µP Cache | µP Cache | µP Cache | µP Cache

Circuit Board

Router

SIA Semiconductor Roadmap

- Generalization of Moore's Law
  - Projects many parameters
  - Years through 2016
  - Includes justification
  - Panel of experts
    - known to be wrong at times
  - Size between Albuquerque white and yellow pages



Projected Interconnect Bandwidth

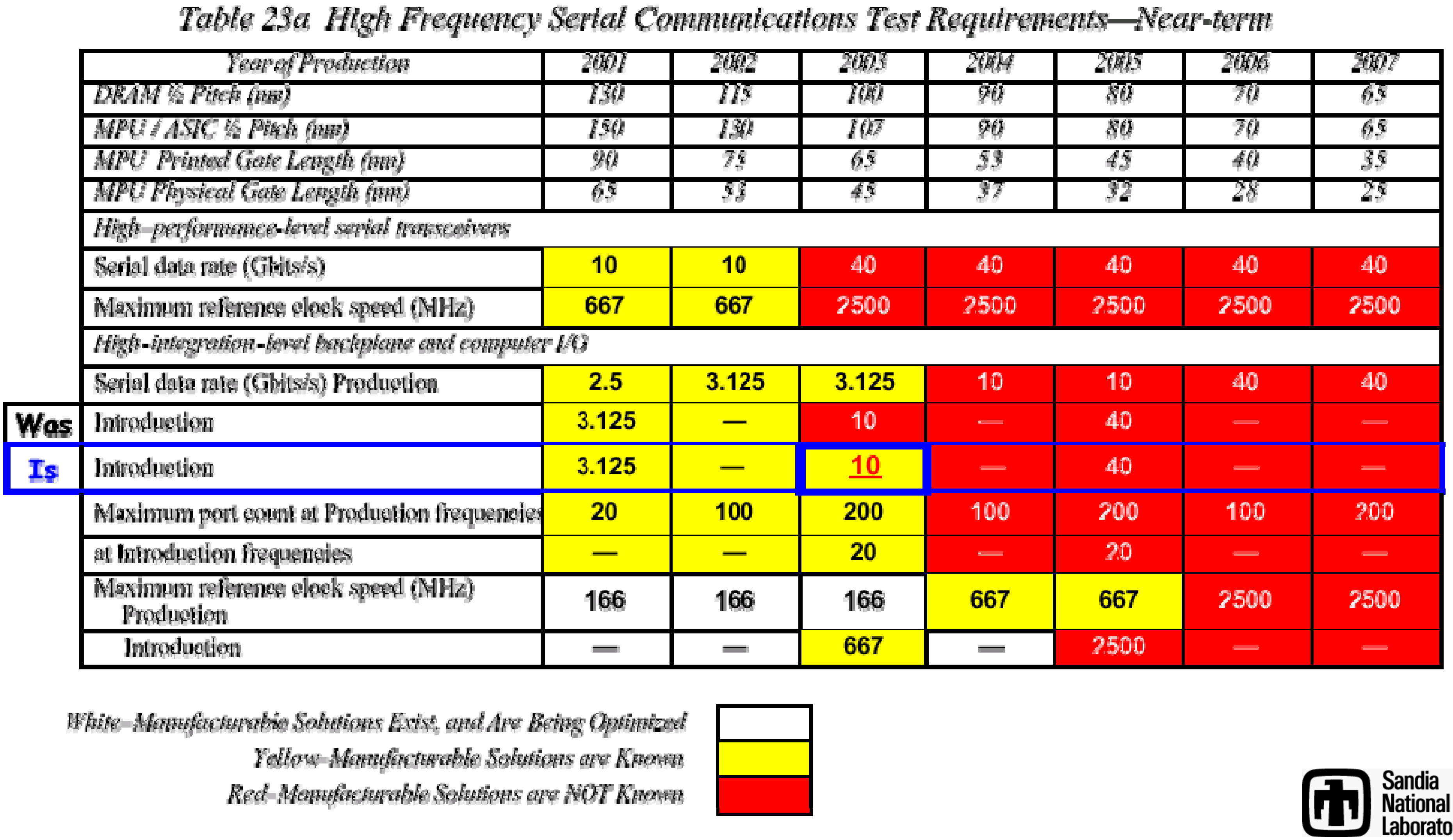


Table 23a High Frequency Serial Communications Test Requirements—Near-term

	Year of Production	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*
	DRAM ½ Pitch (nm)	*130*	*115*	*100*	*90*	*80*	*70*	*65*
	MPU / ASIC ½ Pitch (nm)	*150*	*130*	*107*	*90*	*80*	*70*	*65*
	MPU Printed Gate Length (nm)	*90*	*75*	*65*	*53*	*45*	*40*	*35*
	MPU Physical Gate Length (nm)	*65*	*53*	*45*	*37*	*32*	*28*	*25*
	High–performance-level serial transceivers							
	Serial data rate (Gbits/s)	10	10	40	40	40	40	40
	Maximum reference clock speed (MHz)	667	667	2500	2500	2500	2500	2500
	High-integration-level backplane and computer I/O							
	Serial data rate (Gbits/s) Production	2.5	3.125	3.125	10	10	40	40
Was	Introduction	3.125	—	10	—	40	—	—
Is	Introduction	3.125	—	10	—	40	—	—
	Maximum port count at Production frequenceies	20	100	200	100	200	100	200
	at Introduction frequencies	—	—	20	—	20	—	—
	Maximum reference clock speed (MHz) Production	166	166	166	667	667	2500	2500
	Introduction	—	—	667	—	2500	—	—

White–Manufacturable Solutions Exist, and Are Being Optimized	⬜
Yellow–Manufacturable Solutions are Known	🟨
Red–Manufacturable Solutions are NOT Known	🟥

Sandia National Laboratories

Pin Count

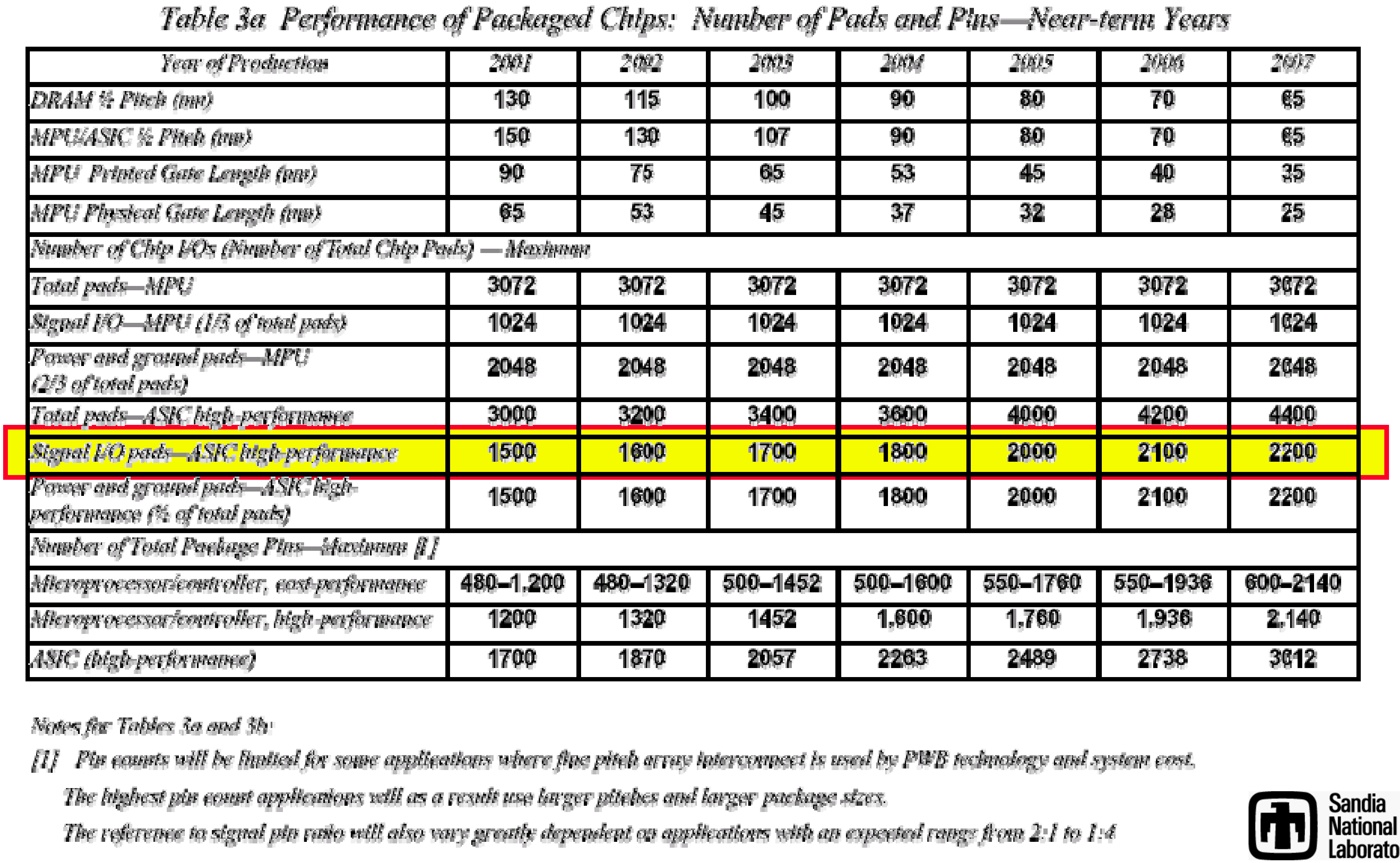


Table 3a Performance of Packaged Chips: Number of Pads and Pins—Near-term Years

Year of Production	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*
DRAM ½ Pitch (nm)	130	115	100	90	80	70	65
MPU/ASIC ½ Pitch (nm)	150	130	107	90	80	70	65
MPU Printed Gate Length (nm)	90	75	65	53	45	40	35
MPU Physical Gate Length (nm)	65	53	45	37	32	28	25
Number of Chip I/Os (Number of Total Chip Pads) — Maximum							
Total pads—MPU	3072	3072	3072	3072	3072	3072	3072
Signal I/O—MPU (1/3 of total pads)	1024	1024	1024	1024	1024	1024	1024
Power and ground pads—MPU (2/3 of total pads)	2048	2048	2048	2048	2048	2048	2048
Total pads—ASIC high-performance	3000	3200	3400	3600	4000	4200	4400
Signal I/O pads—ASIC high-performance	1500	1600	1700	1800	2000	2100	2200
Power and ground pads—ASIC high-performance (½ of total pads)	1500	1600	1700	1800	2000	2100	2200
Number of Total Package Pins—Maximum [1]							
Microprocessor/controller, cost-performance	480–1,200	480–1320	500–1452	500–1600	550–1760	550–1936	600–2140
Microprocessor/controller, high-performance	1200	1320	1452	1,600	1,760	1,936	2,140
ASIC (high-performance)	1700	1870	2057	2263	2489	2738	3012

Notes for Tables 3a and 3b:

[1] Pin counts will be limited for some applications where fine pitch array interconnect is used by PWB technology and system cost.

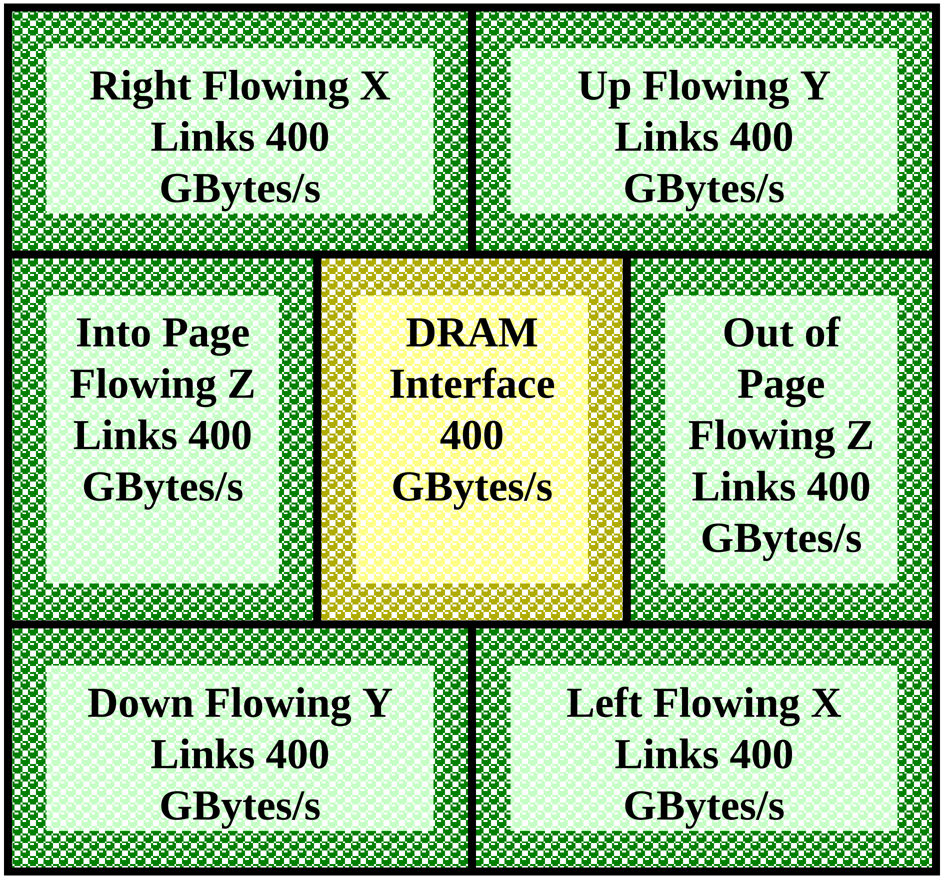
The highest pin count applications will as a result use larger pitches and larger package sizes.

The reference to signal pin ratio will also vary greatly dependent on applications with an expected range from 2:1 to 1:4

Sandia National Laboratories

Chip Interconnect

- Bandwidth ought to be OK for next generation
  - Processor-Memory
  - Processor-Interconnect
- Remarkably unclear that COTS chips will exploit potential
  - This is a risk factor
- Diagram shows approximate proposed chip interconnect budget
- Bumps represent off-chip connections



Right Flowing X Links 400 GBytes/s

Up Flowing Y Links 400 GBytes/s

Into Page Flowing Z Links 400 GBytes/s

DRAM Interface 400 GBytes/s

Out of Page Flowing Z Links 400 GBytes/s

Down Flowing Y Links 400 GBytes/s

Left Flowing X Links 400 GBytes/s

Key Issue: Latency

- Of all semiconductor parameters, the speed of light (c) has fallen behind Moore's Law more than all others
  - – c has not changed measurably in the last 30 years
  - – c is decreasing exponentially with time when measured in distance traveled per clock period

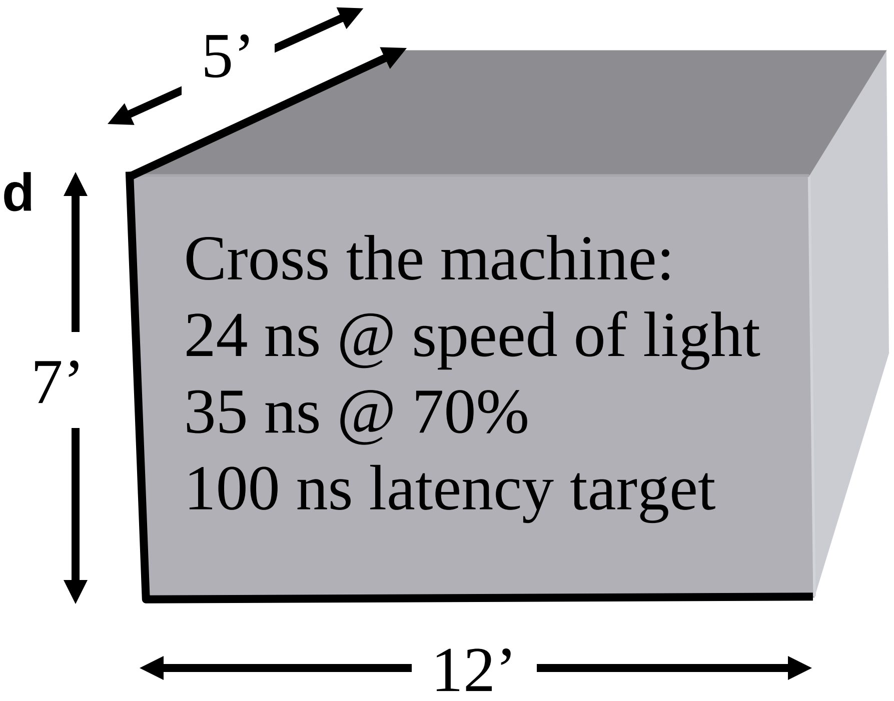
Joke

- Options
  - – Machine is so large that 100 ns latency is not possible due to c.
    - Relax constraint
    - Not a good options because application scalability unknown with unbalanced latency
  - – Cut size of machine
    - 3D packaging

Sandia National Laboratories

Cut Size of Machine

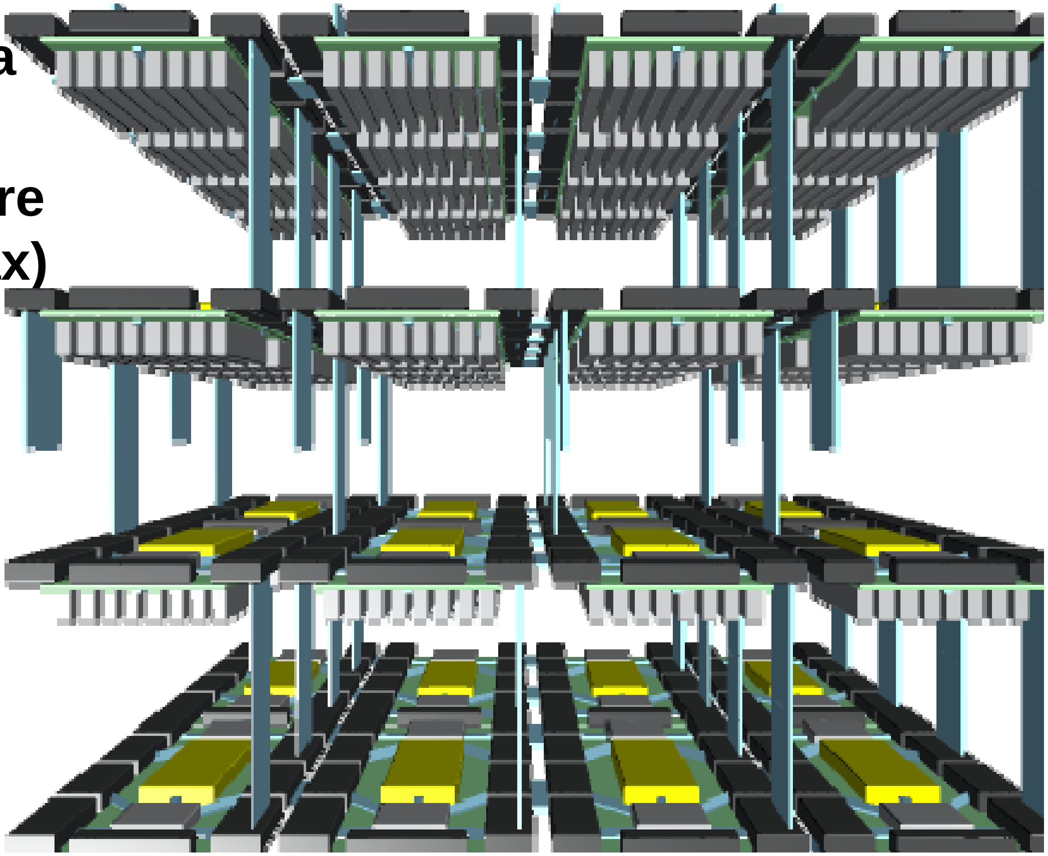
- Water Cooling
  - 10,368 nodes (16×27×24)
  - Diagram to right would only be possible with water cooling
  - 100 ns latency
- Air Cooling later
- About 10 ns budget for MPI software stack overhead
  - (Need to talk to Barney)



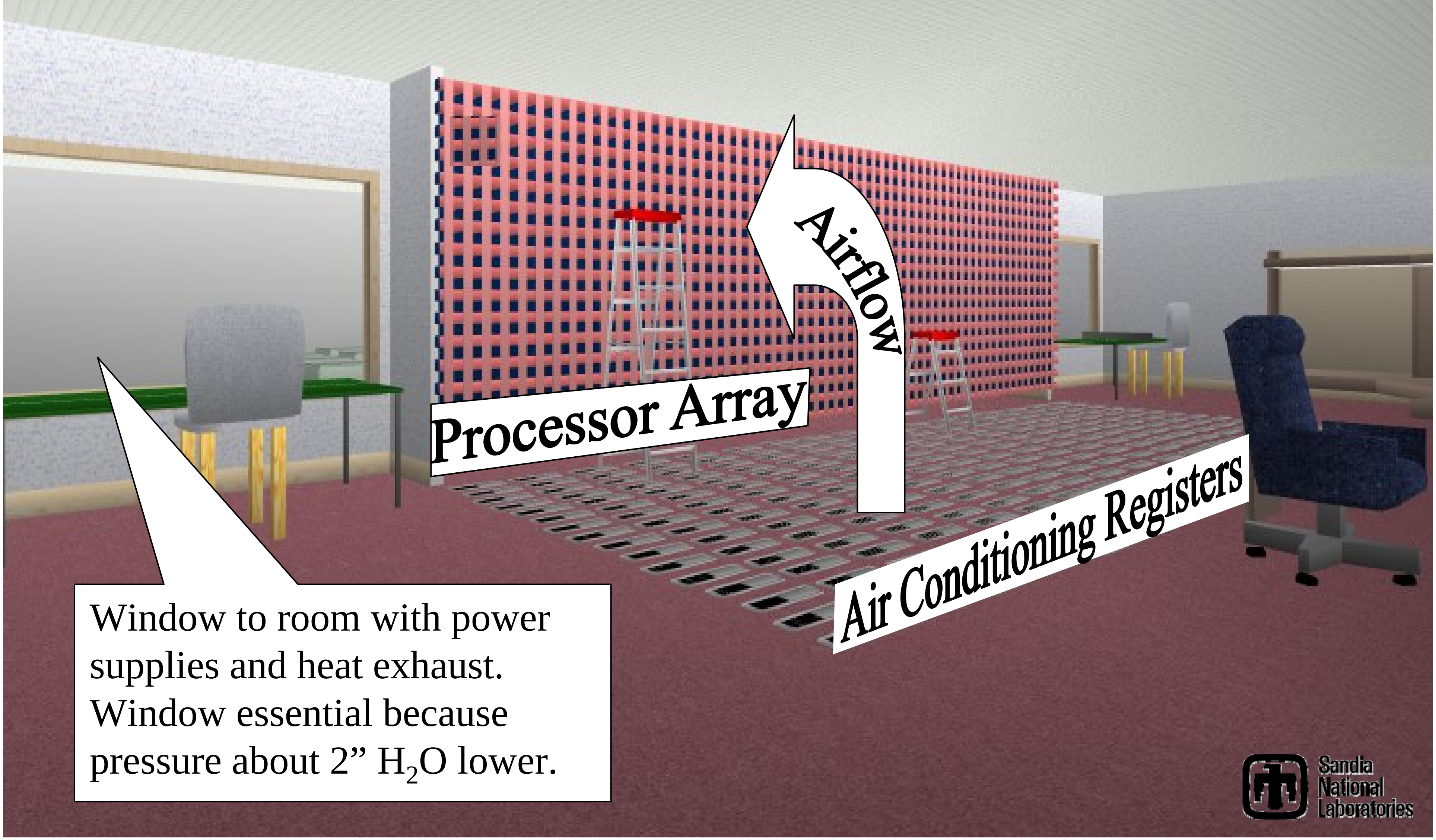
Cross the machine:
24 ns @ speed of light
35 ns @ 70%
100 ns latency target

Homogeneous Packaging

- Entire supercomputer is a single structure
- All mesh network wires are of constant length (8” max)
- Air flows front to back
  - – General approach will work for liquid cooling as well



Air-Cooled Packaging



Window to room with power supplies and heat exhaust. Window essential because pressure about 2” H_2O lower.

Is A Mesh A Good Topology?

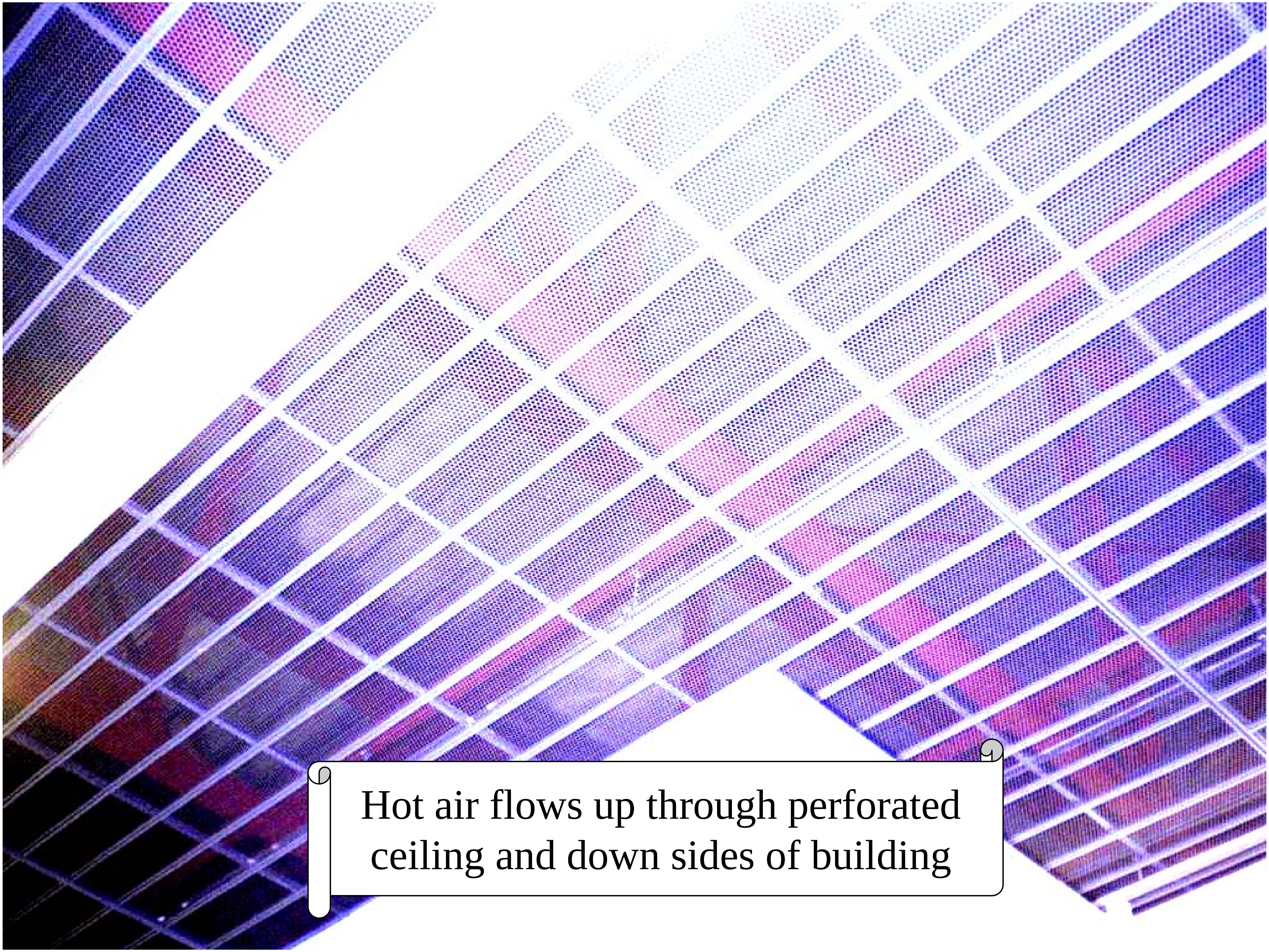
- Mathematicians
  - – Delay related to number of “hops” or network diameter
  - – Not relevant
- Physicist
  - – Delay is distance traveled/c

- Speed of propagation in proposed mesh is c divided by
  - – ÷ √3 Cartesian motion
  - – × propagation velocity in a transmission line (.7)
  - – ÷ curvature of wire (2)
  - – + router delay (1 ns/hop)
- Still within a constant factor of optimal!

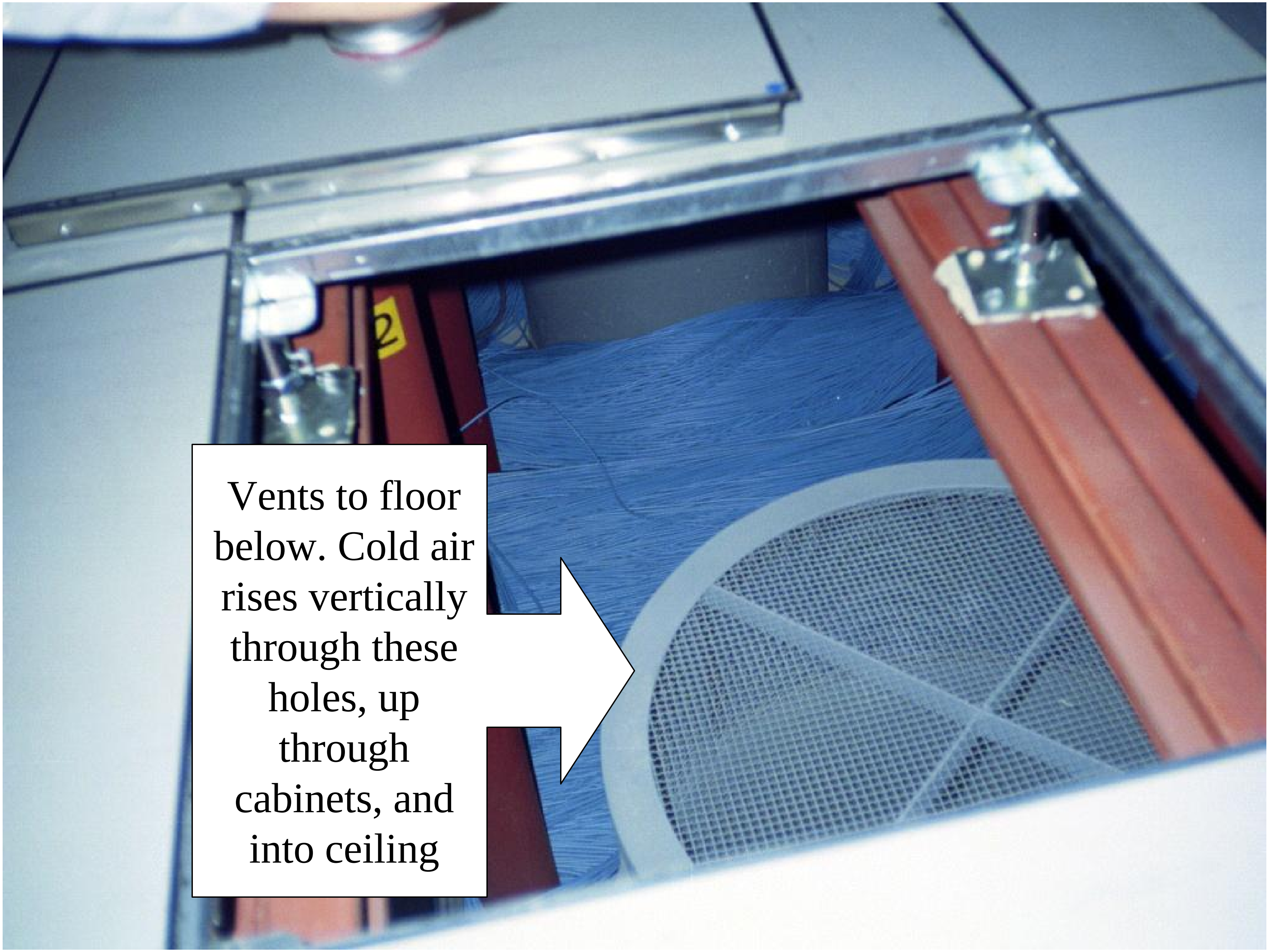
Has Anybody Made a 3d Machine?

- All sorts of university prototypes
- Idea would be more credible if there were a successful example
- Let’s see…

P



Hot air flows up through perforated ceiling and down sides of building



Vents to floor below. Cold air rises vertically through these holes, up through cabinets, and into ceiling

Reliability

- Red Storm
  - Separate RAS network (2500 Unix processors & LAN)
  - Central point of information collection and control of entire machine
  - Capable of halting running machine, permitting deconfiguration of a faulty node, restart
  - Red Storm uptime specs: 50 & 100 MTBF/MTBI
    - If your PC had this MTBF, you’d take it back to Frys

Reliability Forward

- Cosmic Rays
  - These will be a problem in the next generation. COTS microprocessors have some tens of thousands of unprotected flip flops. This impacts an ASCI size machine on 6 month timeframe

- FIT Rate
  - Manufacturers can give (under NDA) fit rate for components when used in a commercial environment
  - Predicts ~1 hour MTBF
  - However, machine rooms are temperature controlled and power is not cycled
  - Actual "weeks" MTBF

Conclusions

- A 1000 TFLOPS successor to Red Storm is an engineering challenge
- Risk factors
  - SMP nodes
  - Memory bandwidth
  - Need new network interface
- Will 10 PFLOPS?
  - Scaling, speed of light, memory wall, threads
- COTS vs. Custom
  - Unknown which will "win"
  - Prepare to deal with both
  - Have capability on all key hardware issues
    - HDL
    - FPGA
- Shared address space?
  - 100 ns network makes this possible