SILICON GRAPHICS | The Source of Innovation and Discovery™

High-Performance Computing at SGI and the Status of Climate and Weather Codes on the SGI Altix Gerardo Cisneros, Ph.D. Scientist

Sgi

C2004 Silicon Graphics, Inc. All rights reserved. Silicon Graphics, SGI, IRIX, Origin, Onyx, Onyx2, IRIS, Altix, InfiniteReality, Challenge, Reality Center, Geometry Engine, ImageVision Library, OpenGL, XFS, the SGI logo and the SGI cube are registered trademarks and CXFS, Onyx4, InfinitePerformance, IRIS GL, Power Series, Personal IRIS, Power Challenge, NUMAflex, REACT, Open Inventor, OpenGL Performer, OpenGL, Optimizer, OpenGL Volumizer, OpenGL Shader, OpenGL Multipipe, OpenGL Vizserver, SkyWriter, RealityEngine, SGI ProPack, Performance Co-Pilot, SGI Advanced Linux, UltimateVision and The Source of Innovation and Discovery are trademarks of Silicon Graphics, Inc., in the U.S. and/or other countries worldwide. Linux is a registered trademark of Linus Torvalds in several countries, used with permission by Silicon Graphics, Inc. Intel and Itanium are registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries. Red Hat and all Red Hat-based trademarks are trademarks or registered trademarks of Red Hat, Inc. in the United States and other countries. Linux penguin logo created by Larry Ewing. All other trademarks mentioned herein are the property of their respective owners. (04/04)

Overview

- Company focus
- SGI Altix: present and future
- Performance of NWS codes
- Conclusions

Silicon Graphics

Providing the Industry's Highest-Performing Compute, Storage and Visualization Products

Exclusively focused on the technical computing market
Technology is designed to enable the most significant scientific and creative breakthroughs of the 21st century
Products and services are mission critical to government and defense, science and research, manufacturing, energy and media industries



9/9/2004 Slide 4

Images courtesy of SCI Institute, University of Utah, Navy Rehearsal TOPSCENE Program, Magic Earth LLC, and WETA Digital

Strategic Focus Areas

High Performance Computing



High-performance NUMAflex[™] architecture

delivers unprecedented flexibility and performance.

Storage



Advanced Visualization



CXFS[™] shared file system allows transparent, heterogeneous file access everywhere, without copying data.

Onyx4[™] allows users to combine multiple industry-standard graphics cards in a highbandwidth, low-latency architecture, for cost-effective high performance visualization.

sgi

Architecture Designed for HPC; Choice of Deployments

NUMAflex™ Global Shared-Memory Architecture

Balanced, scalable performance Operating environment optimized for HPC Low-latency memory access Easily Deployable

MIPS[®] and **IRIX[®]**



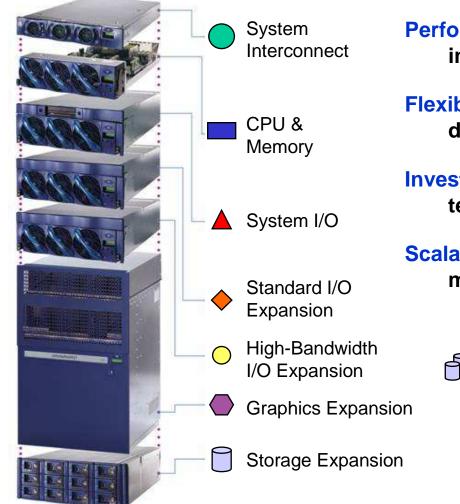
SGI[®] Origin[®] Family

Intel[®] Itanium[®] 2 and Linux[®]



SGI[®] Altix[®] Family

Modular SGI[®] NUMAflex™ Architecture

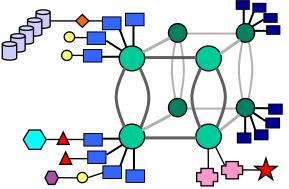


Performance: High-bandwidth interconnect with very low latency

Flexibility: Tailored configurations for different dimensions of scalability

Investment protection: Add new technologies as they evolve

Scalability: No central bus or switch; just modules and NUMAlink[™] cables

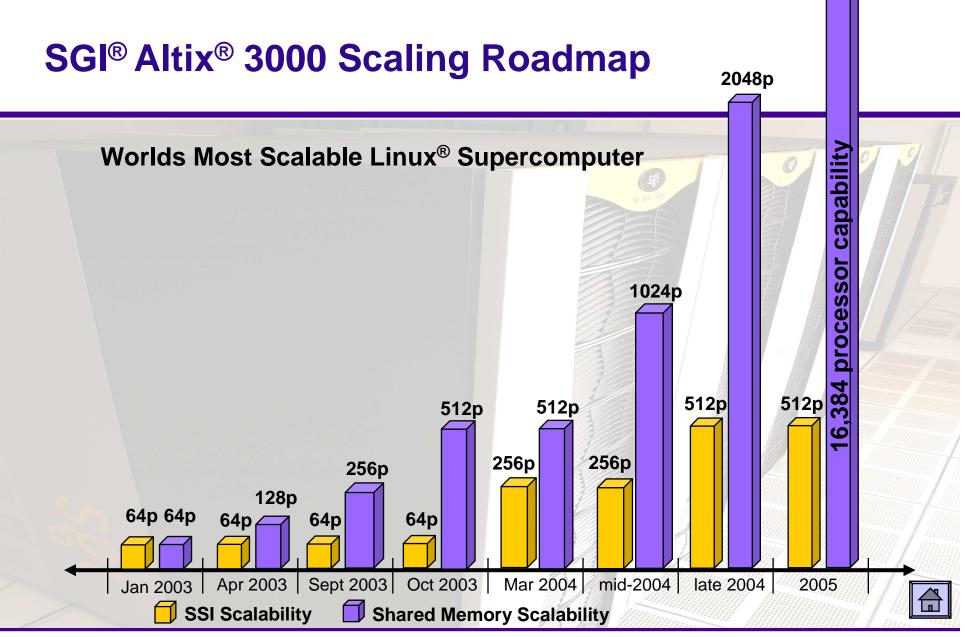


sgi

SGI Family of Scalable Linux[®] Solutions

SGI[®] Altix[®] 350 Servers and Clusters SGI[®] Altix[®] 3000 Servers and Superclusters



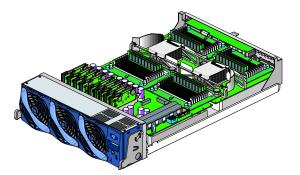


9/9/2004 Slide 9

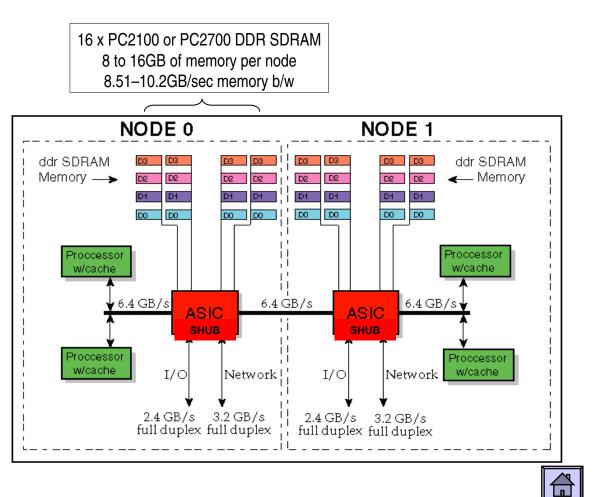
This slide contains forward-looking statements. The results and forecasts as stated may vary. Other risks and uncertainties relating to this slide may be found in the "Safe-Harbor" statement at the beginning of this presentation.



SGI® Altix® 3000 C-Brick Detail

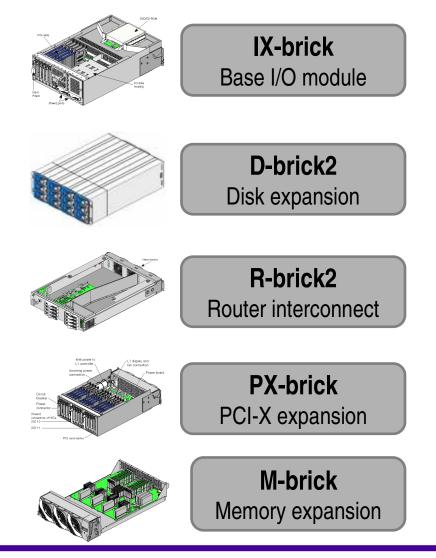


- 4x Intel[®] Itanium[®] 2 processors
- 2 processor per 6.4GB/sec frontside bus
- 4–64GB memory C-brick
- SHUB memory controller 8.51–10.2GB/sec memory bandwidth (varies with memory speed 133 MHz vs. 166 MHz)
- 6.4GB/sec aggregate interconnect bandwidth
- 4.8GB/sec aggregate I/O
 bandwidth

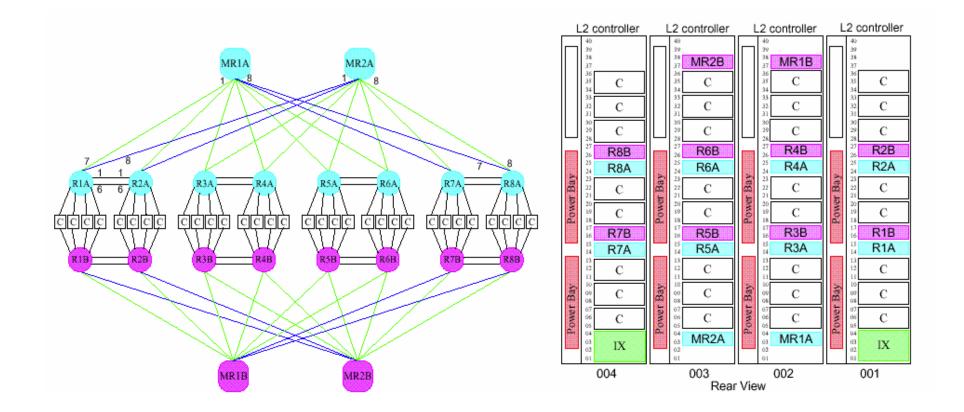


SQ

Other Altix Bricks



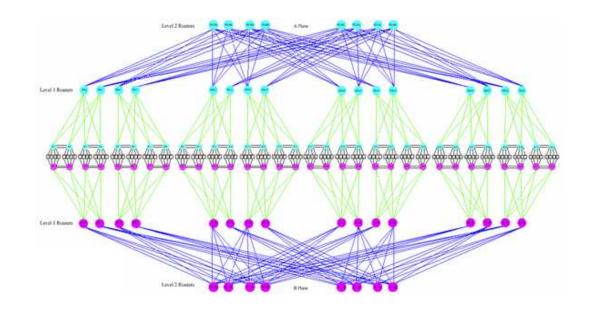
Altix 3700 — 128 processors



9/9/2004 Slide 12

sgi

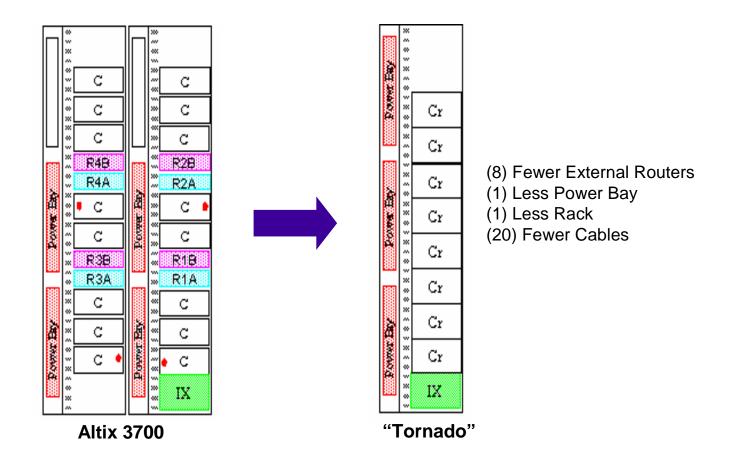
Altix 3700 — 512 processors



2 Controller	L	2 Controller	- 20	Controller	14	Controller	14	Controller	12	Controller	1	530	ontroller	<u></u>	2.0	0.000
ax.		RB23A	n	RB22A RB6A		RC3A		RC9A RC7A	ini	RB30A RB14A	Π		RB31A RB15A	n		ix
C		f C		C		F C		i c		c			¢.			C
C		C		C		C		C		c		В	c		13	C
C		C C				C		C		C		B	0		19	c
R22A	Ľ	R24A	Ľ	C REA	2	RMA	Ш	RIAA	H	RIEA	Ľ	12	RINA	Ľ	15	832
R220		R248		Rea		RM		R14B		R168			R306			1132
C	E	c		C		C		c		C	-	1	C.	4	T	C
C		C		C		C		- C		C	Del.	B	C	and a		C
R21A		R23A	1	RSA		R7A		RIBA		R15A	h	2	FI29A		12	R31
R210		R238		ASB		R78		R138		R150		1	R291		3	Rat
C		C		C		C		C		c		13	0			¢.
c		c	in the	c		C	1	C.	2	C	1	1	c	Mar	E	C
C	H	c	П	C		C	2	E C		C	1	E	C.		E	C
		RE118		RETOR		RC58		RCSE		R8148			RB158	K.		IX
								RC78		48308	-		RB318			
011 Controller		012	12	Controller	12	Corecter	12	015 Controller	12	D10		20	017		20	
011		012	12	013 Corector	12	054	12	015	12	-016	L	20	017		20	
011	ľ	012 2 Controller 5 RB1BA	F	013 Controller	LI LI	Core other	ň	Controller		010 Contruiter RB27A	Б	20	017 ontroller RB26A	F	20	
011 Controller	ľ	012 Cormoler RB18A RB2A	ň	013 Corector		Os4	ň	015 Controller		D16 Controller RB27A RB11A	ĥ	2011111	017 RB26A RB10A	F	201111	entros EX
011 Controller UX		012 Comolor CRB1BA RB2A	ľ	013 Controller RB19A RB3A C		Corecoller E RC4A RC2A	12	015 Controller RC8A RC6A		Contrulier RB27A RB11A C	ĥ	Constraint of	017 RB26A RB10A C		Guntari	EX C
011 Controller IX C		012 Controller RB1BA RB2A C		Controller RB19A RB3A C		Corecter RD4A RD4A RD2A C	12	Controller RCBA RCBA C		D16 Controller RB27A RB11A C C		Constanting of the second	017 RB26A RB10A C C		Circle 1 and	c c c
011 Controller C C		012 Commoler C RB1BA		Controller RB19A RB3A C C		Controller RC4A RC2A C C		Controller RCBA BCGA C C		D16 R827A R811A C C		Construction of the	017 RB26A RB10A C C		Circle in the second	C C C
011 Controller 6X C		012 Controller RB1BA RB2A C		Controller RB19A RB3A C		Corecter RD4A RD4A RD2A C		Controller RCBA RCBA C		D16 Controller RB27A RB11A C C		Construction of a	017 RB26A RB10A C C		Constantination of the later	C C C R28
011 IX C C R18A		012 Commoler RB1BA FB2A C C R20A		Controller RB19A RB19A C C C C RDA	Bog []]	Controller Ricka Ricka C C C Ricka Ricka Ricka Ricka		Corboiler RCBA RCBA C C R10A		D16 RB27A RB11A C C C R12A	Ma I	Garage State State State	017 RB26A RB10A C C C R2EA	Hare a second se	Constant and a state of the sta	C C C R28
011 Controller IX C C C R16A R160	the flag	012 012 Corection RB18A FB2A C C R20A R20A	On or Hay	Controller RB19A RB3M C C C R2A R2B	the first state of the state of	Controller RIC4A RIC2A C C C C R4A R4H	wet they	Controller RCBA RCGA C C R10A R10A		D16 RB27A RB11A C C C R12A R12A R12B	ello.	Constrainted at the loss	017 entroller RB256A RB10A C C C R26A R266	of Bay	Construction and the state	C C C R28 R29
OTT Controller C C C R16A R160 C	Plane Ruy	Controller Controller RB18A C C C R20A R20A C C C R20A R20A C C C R18A	Plane Bay	013 Controlar RB19A RB3A C C C C R2A R2B C C C R2A R2B C C R2A R2B	Poset Boy	Commer RD4A RD2A C C C C R4A Rd8 Rd8 C C C R4A Rd8 Rd8 C C C R4A	at the	O15 Controller RCGA RCGA C C R10A R568 C C R3A		D16 Controller RB27A RB11A C C C C R12A R12B C C R11A	Ma I	Gundrefenfreten tetreten	017 RB26A RB10A C C C R26A R260 C C R25A	Means that	Construction and the state	C C C R28 R29 C C C
011 Controller IX C C C C C R18A R18B R17A R17B	how hay	012 012 Comoler RBSBA RBZA C C C R2DA R2DA R2DA R2DA R2DA R190	Rease Ray []	D13 CONCENT RB19A RB19A C C C C C R2A R29 C C R2A R19 C C C R2A R19	Fost Bay	Commerce RDAA RD2A C C C C RAA RD3 RD3 RD3 RD3 RD3 RD3 RD3 RD3 RD3 RD3	wet they	O15 Controller RCGA RCGA C C R10A R568 C R3A R669		016 RB27A RB11A C C C R12A R12B C C C	ello.	Construction of the second second	017 RB26A RB10A C C C R26A R260 C R26A R260 C C R25A R250	of Bay	Construction of the other of the	C C C R27 R27 R27 R27
011 Controller IX C C C C C R180A R180B C C C R180A	Now Boy	Controller Controller RB18A C C C R20A R20A C C C R20A R20A C C C R18A	Water Bay	013 Controlar RB19A RB3A C C C C R2A R2B C C C R2A R2B C C R2A R2B	Four Day	Commer RD4A RD2A C C C C R4A Rd8 Rd8 C C C R4A Rd8 Rd8 C C C R4A	wet they	O15 Controller RCGA RCGA C C R10A R568 C C R3A		D16 Controller RB27A RB11A C C C C R12A R12B C C R11A	Prover Muy	Constrainted as I device a state	017 RB26A RB10A C C C R26A R260 C C R25A	of Bay	Construction of the owner own	C C C R28 R28 C C C R27
011 Controller IX C C C C C R18A R18B R17A R17B	Boy Now Boy D	012 012 Comoler RBSBA RBZA C C C R2DA R2DA R2DA R2DA R2DA R190	Tap Nume Hay	D13 CONCENT RB19A RB19A C C C C C R2A R29 C C R2A R19 C C C R2A R19	Ber Powerber	Commerce RDAA RD2A C C C C RAA RD3 RD3 RD3 RD3 RD3 RD3 RD3 RD3 RD3 RD3	wet they	O15 Controller RCGA RCGA C C R10A R568 C R3A R669		D16 Controller R0277A R011A C C C R12A R12D C C R11A R110	ello.	Construction of the second states of the second sec	017 RB26A RB10A C C C R26A R260 C R26A R260 C C R25A R250	of Bay	2 a subving a subving the subving the subving	C C C R27 R27 R27 R27
011 Controller C C C R 198A R 198A C C C R 198A C C C C R 198A C C C C C C C C C C C C C C C C C C C	the Bay Nove Bay	C C C C C C C C C C C C C C C C C C C	trans Raje [Power Bay]	D13 CONCENT RB19A RB3A C C C C C R2A R2B C C R2A R2B C C C C C C C C C C C C C	The Bay Prover Day []	O14 Control of the second sec	wet they	015 Controller RCEA RCEA RCEA C C REEA REEA C REEA REEA		Did Controller RB27A RB11A C C C C C R12A R12A R12A R12A R12A R12A R14A R110 C	Prover Muy	Construction of the second states of the second sec	017 RB26A RB10A C C R24A R260 C R25A R25A R254 R254 R254	of Bay	2 Construction and a first of the second second	1X C C C R28 R28 R29 C C C R27 R27 C C C C C C C C C C C C C C C C C C C
011 Controller C C C R 180 R 180 R 180 R 180 R 170 C C C C C C C C C C C C C C C C C C C	Nuc Sof Nuc So	012 012 Controller RB18A RB2A C C R20A R205 C R204 R205 C C R196 C C	Phone Ray Phone Ray	D13 Controlar RB19A C C C C R2A R2B C C R1A R1B C C C R1A	Prace Bay Power Bay	O14 Controller RICAA RICZA C C C RIAA RAM RAM C C C RIAA RAM RAM C C C C C C C C C C C C C	wet they	015 Controller RCEA RCEA RCEA C C C R10A R506 C C C R3A R066 C C C C		Did Regional Regional Controller Regional Controller Regional Controller Cont	Prover Muy	Garager and a standard of a	017 REDOGA REDOGA C C R26A R260 C R26A R260 C R25A R256 C C C C C C C C C C C C C C C C C C C	of Bay	Construction and a destruction of the Destruction	C C F28 F28 F27 F27 F27 C C



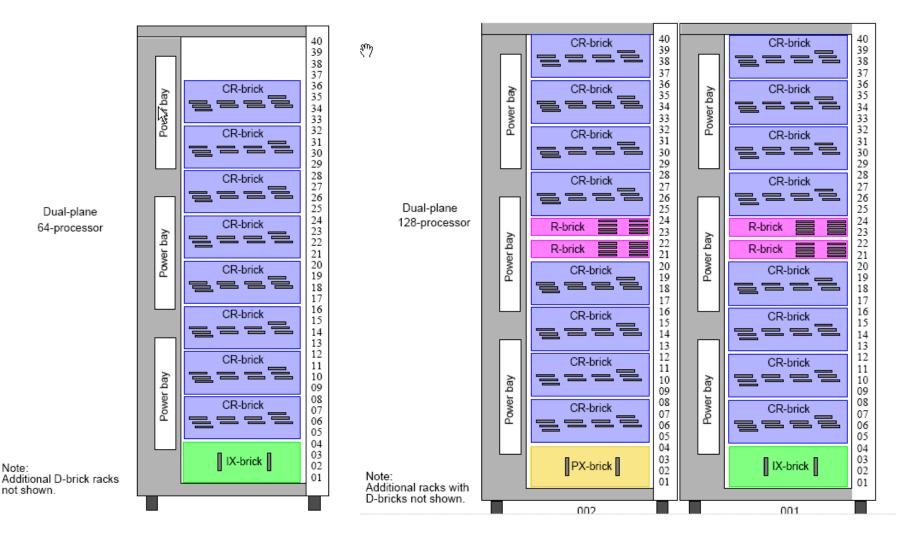
Next generation Altix







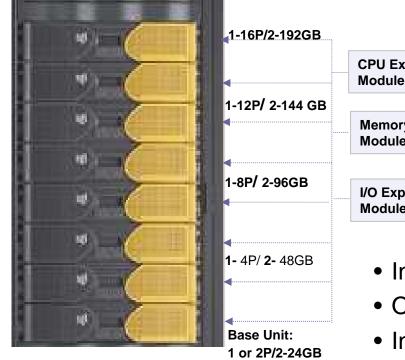
Next generation Altix



9/9/2004 Slide 15

sgi

The Altix[®] 350: "Expand on Demand" Growth Path



CPU Expansion Module

Right-size systems for the ultimate price/performance

- Independently scale CPU, memory, I/O
- One Linux[®] instance to manage
- Investment protection & leverage current assets
- Allocate budget and resources to ongoing needs

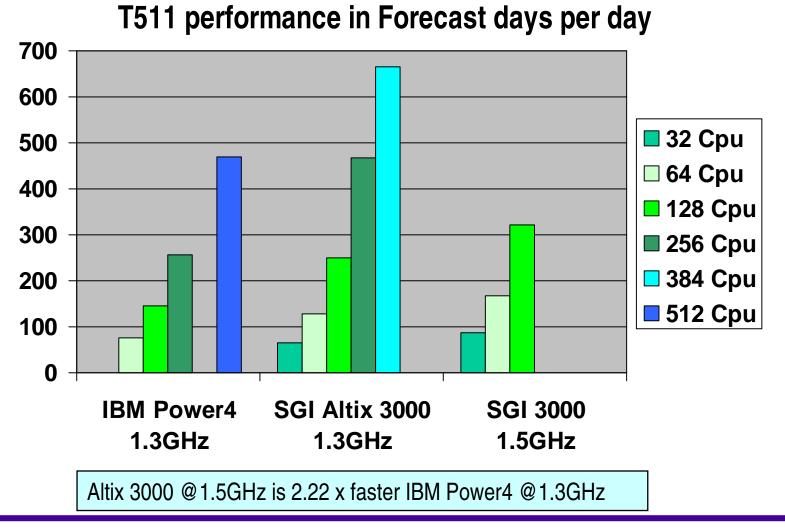


Customers with Altix systems running climate and weather applications

- NASA 10240p ECCO, CAM, CCSM, etc.
- NCSA 1024p WRF
- GFDL 608p (2x256, 1x96) MOM4, CM2.1
- NRL 384p (1x128, 1x256) various
- ORNL 256p CCSM, CAM, POP
- LANL 256p POP, HYCOM
- BAMS 20p MM5 and MAQSIP
- CMMACS 12p Altix350 MOM4
- APAT 8p Altix350 MM5
- Romanian Nat'l Met Admin 2p Altix350 HRM

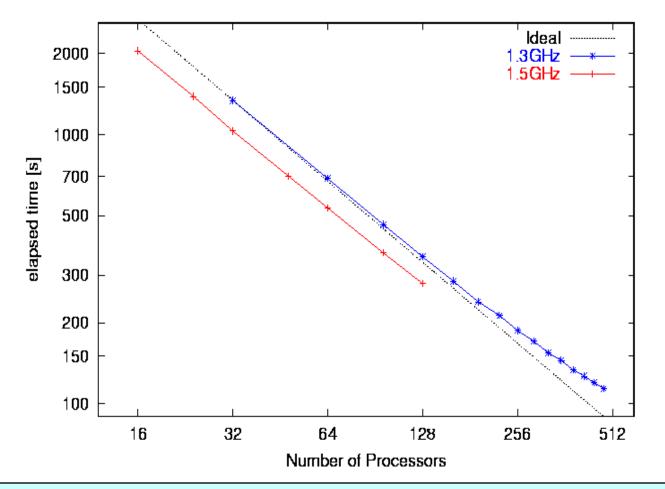


IFS - Performance on SGI Altix 3000





IFS - Scalability on SGI Altix 3000



Itanium2 @1.5GHz is 1.3x faster than Itanium @1.3GHz because of the larger cache and higher clock rate

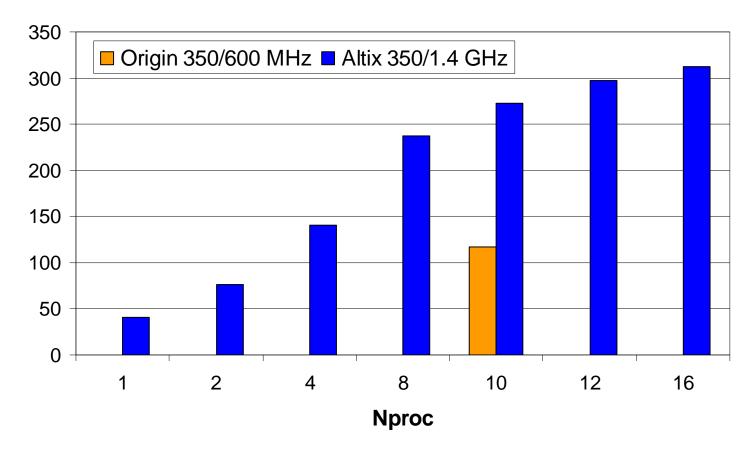
9/9/2004 Slide 19

Source: Roland Richter, SGI, May 2004



HRM Performance on the Altix 350

Simulation speed

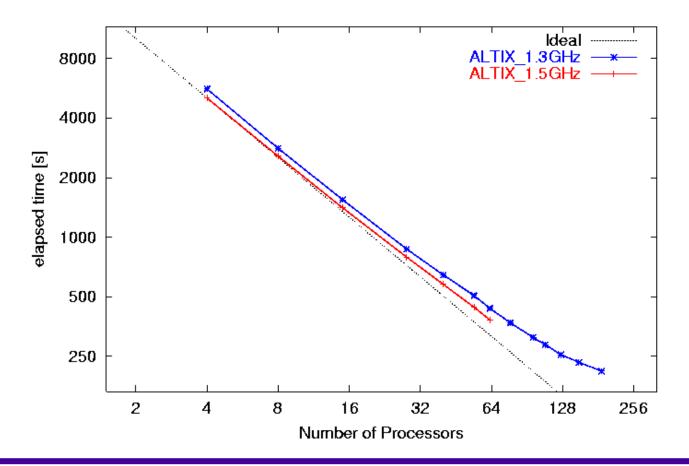


78h forecast in 2340 time steps over a 181x217 horizontal grid with 26 vertical levels

Sg

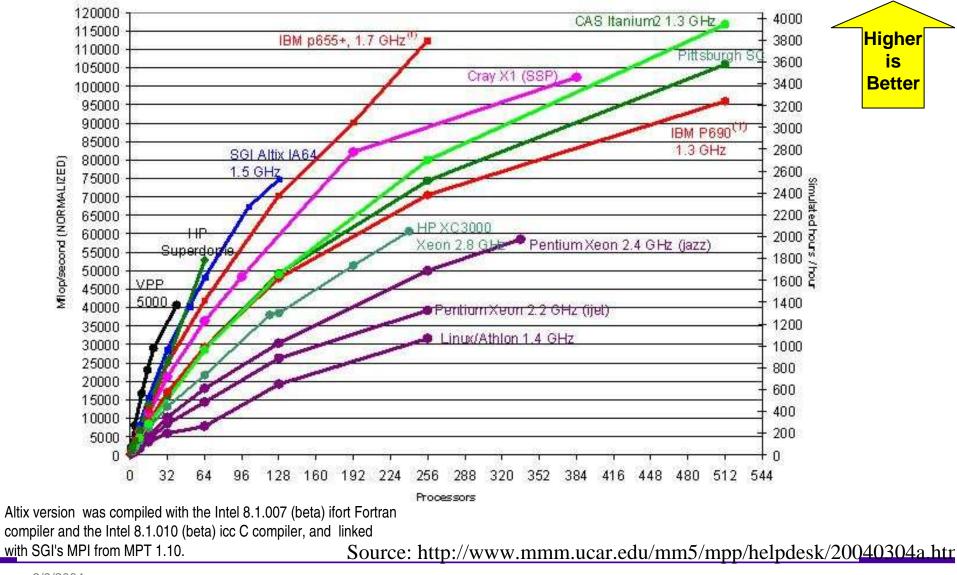
LM-RAPS 2.1 (Optimization using shared Memory)

Scalability on Altix using Intel Compiler and SGI's MPT library



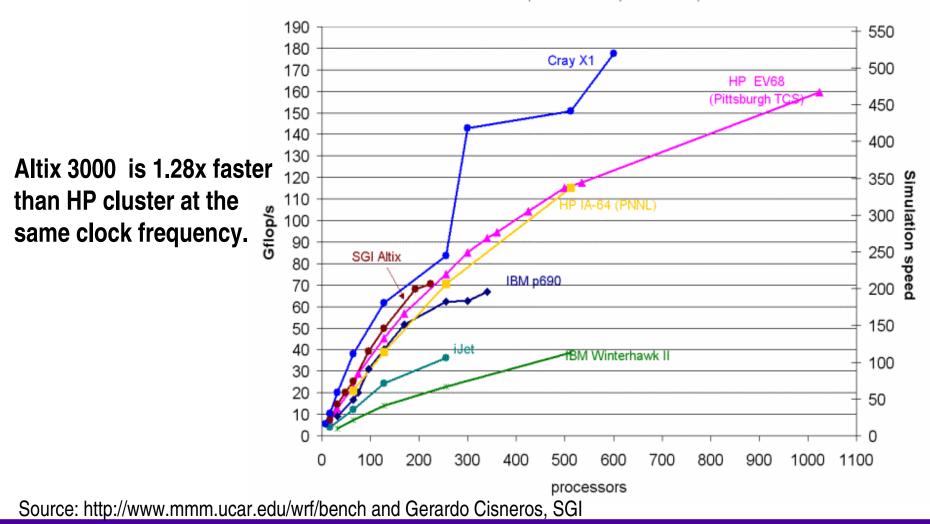
Benchmark case used by Emy (Greece) in 2003 Grid 363x263x45

MM5 3.6.3 - Standard benchmark, 2004



SQ

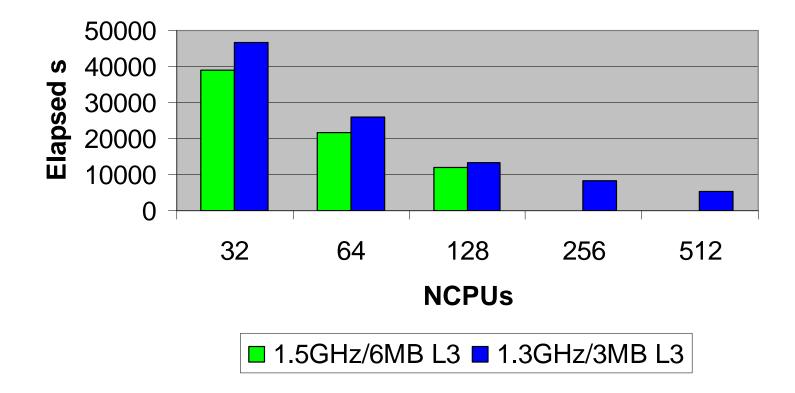
WRF 1.3 - Scalability & Performance on Altix 3000



WRF EM Core, 425x300x35, DX=12km, DT=72s

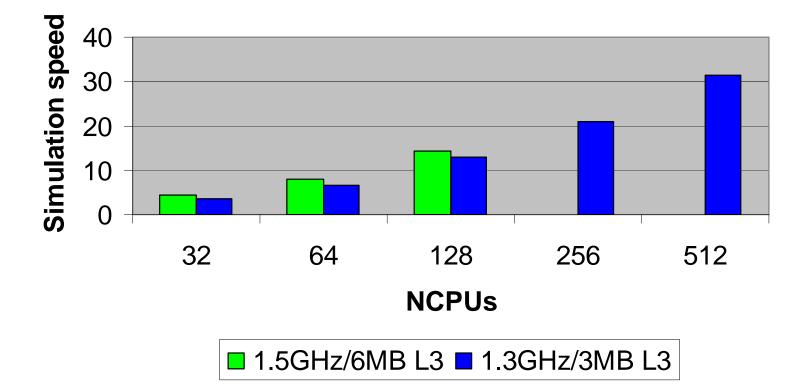
WRF 2.0.2 on a Large Problem

WRF SI 5km CONUS 48h forecast (980x720x37, 5km, 30s)



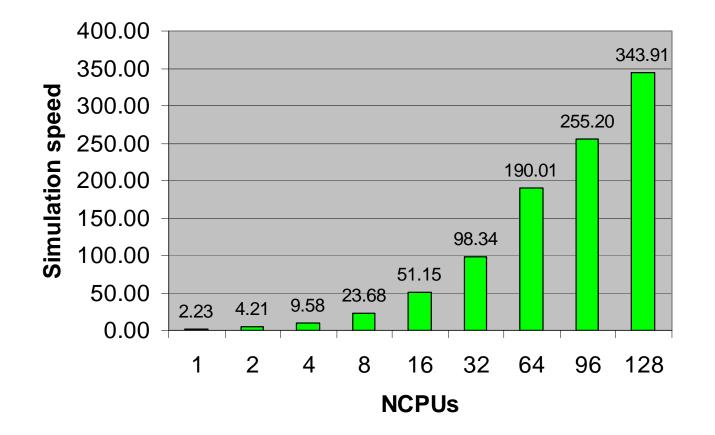
WRF 2.0.2 on a Large Problem

WRF SI 5km CONUS 48h forecast (980x720x37, 5km, 30s)



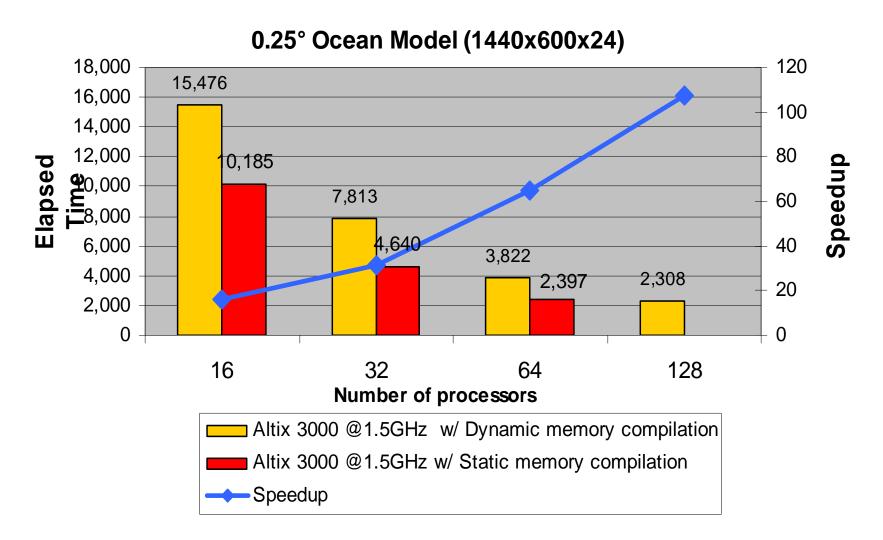
GFS Performance on the Altix 3700 (1.5GHz)

GFS T240L30 (720x360x30) 120h forecast





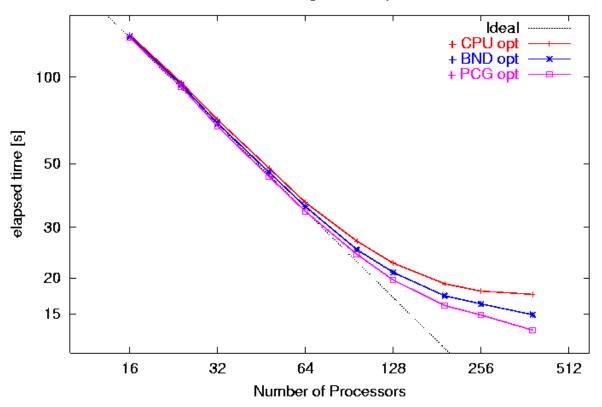
MOM4 - Performance on SGI Altix 3700



Source: Gerardo Cisneros, SGI, May 2004



Scalability on ALTIX using Intel Compiler and SGI's MPT library

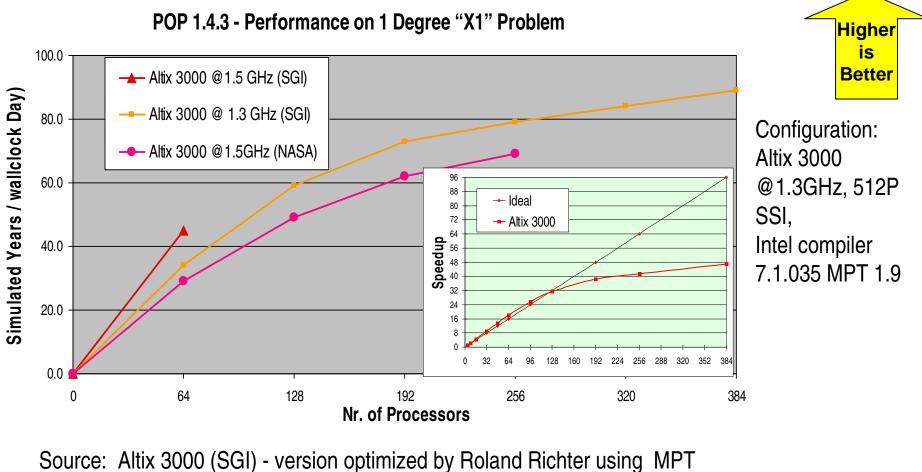


POP Solver scaling on a 5 days simulation

9/9/2004 Slide 28

Increasing Message length and reducing synchronization frequency improved the performance by 20% at 256 CPU

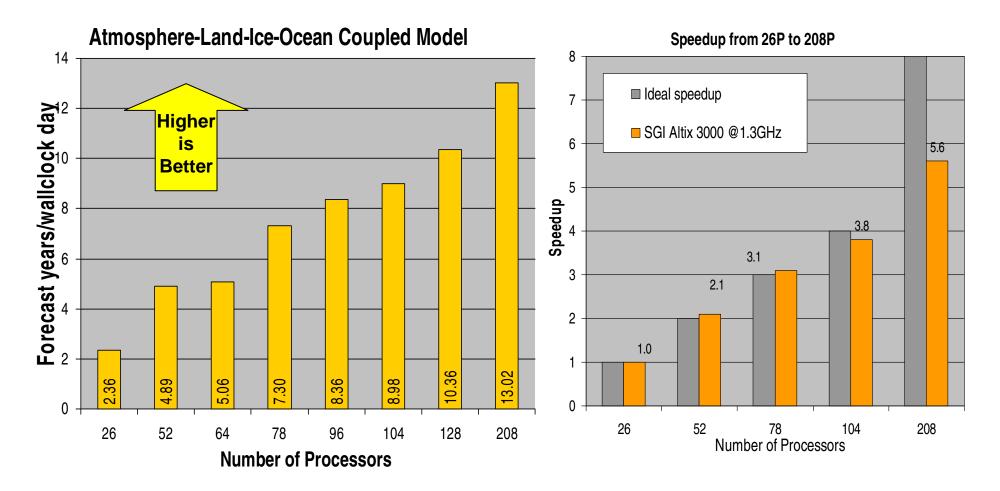
POP 1.4.3 - Performance on SGI Altix 3000 SSI



Altix 3000 (NASA) - Jim Taft's version using MLP



CCSM 3.0 (beta08) Performance on SGI Altix 3000



Run on 512P Altix 3000 @1.3GHz with efc 7.1.035 compiler

Source: Roland Richter, SGI

Conclusions

- The Altix is proving to be an excellent system for weather and climate codes
- Shared memory allows very large models to run
- The compilers are getting better
- The libraries are excellent and still improving
- SGI's software layers on top of Linux help jobs attain their best performance
- SGI engineers have the know-how to make weather and climate codes run well on its customers' SGI systems



