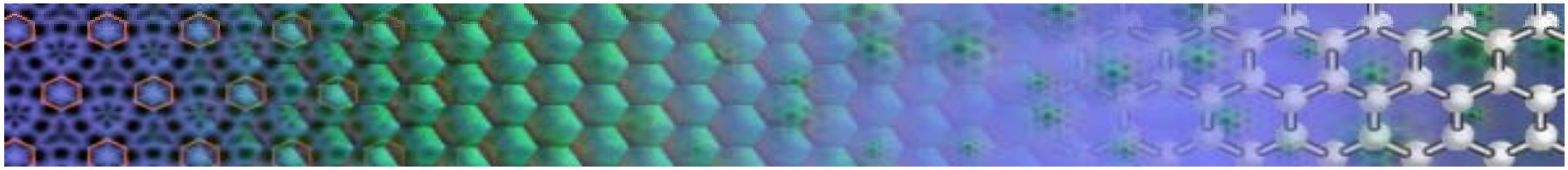
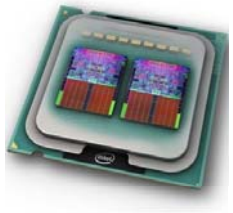


Update on Physical Scalability Sabotaging Performance Gains!

Dallas IEEE Computer Society Meeting
Friday Jan 21, 2011

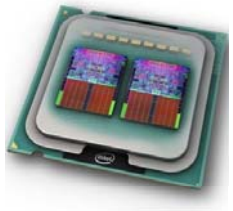
Douglas J. Matzke, Ph.D.
IEEE Senior Member
matzke@IEEE.org

Abstract



In September 1997, Dr. Matzke wrote the lead off paper entitled "Will Physical Scalability Sabotage Performance Gains?" for the special issue of Computer Magazine on "Billion Transistor Computer". This paper is now required reading for most computer architecture courses in the world and cited by 257 other papers. The prediction in that paper was architectures would become more fine grain due to wire scaling and most likely the billion transistor computer would be a multiple CPU machine. This paper will give an update on this prediction and talk about other trends in the architecture and device arena, including multi core cpus, hybrid core machines, Memristors and quantum computing trends.

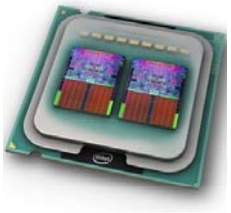
Introduction and Outline

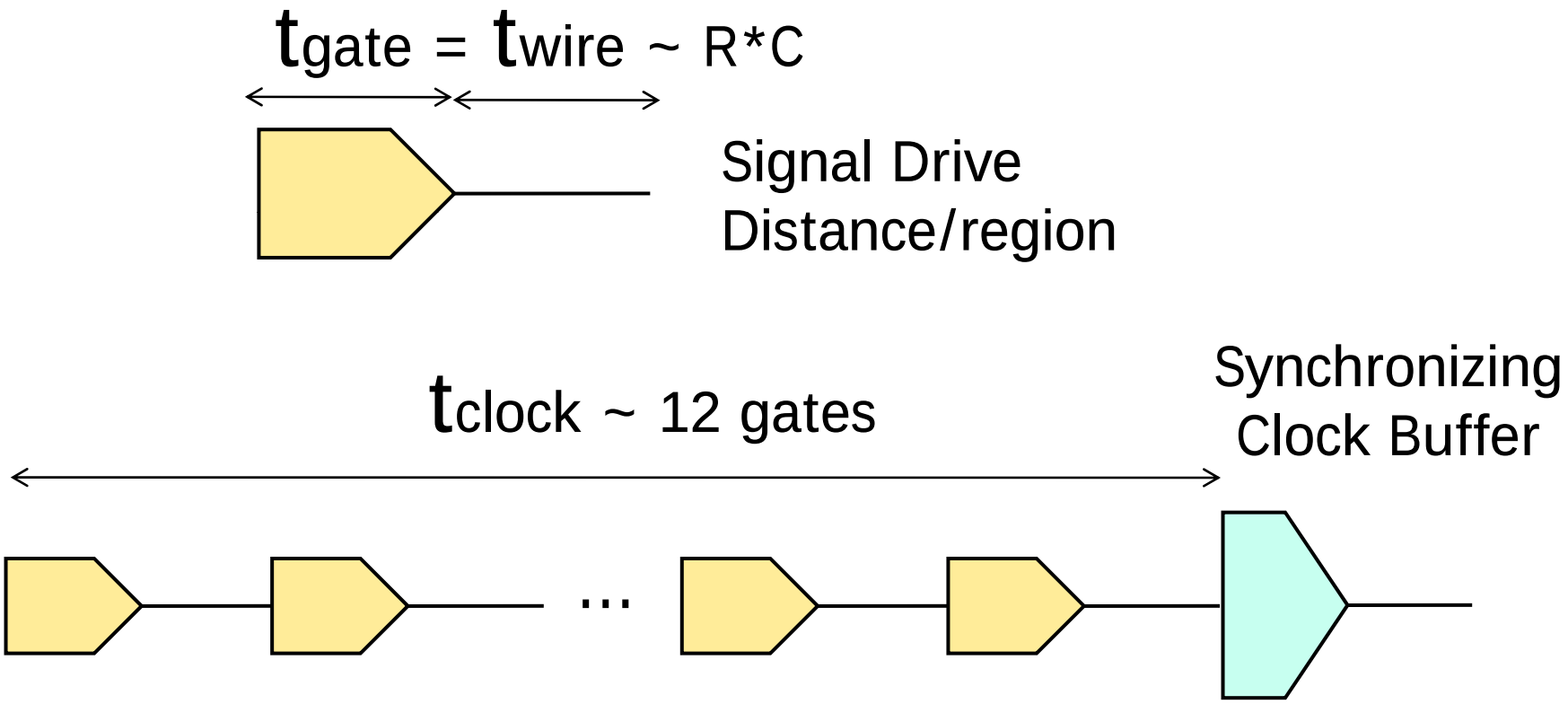


Topics in Presentation

- Review of Wire Scaling Prediction
- Billion Transistor computers
- Current Multi-core processors – Core Wars
- Process Trends and Intel roadmap
- Limits of semiconductor/computer scaling
- Design Trends
- Memristor Fundamentals
- Scaling predictions
- Summary

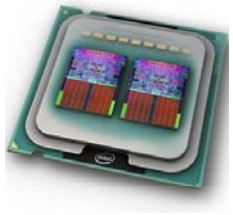
Wire Scaling Prediction 1997





Assumption:
25 simple gate delays per clock or 12 drive distance

Non-scaling Drive Distance

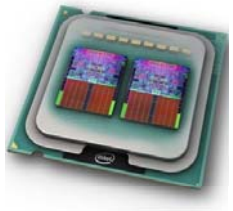


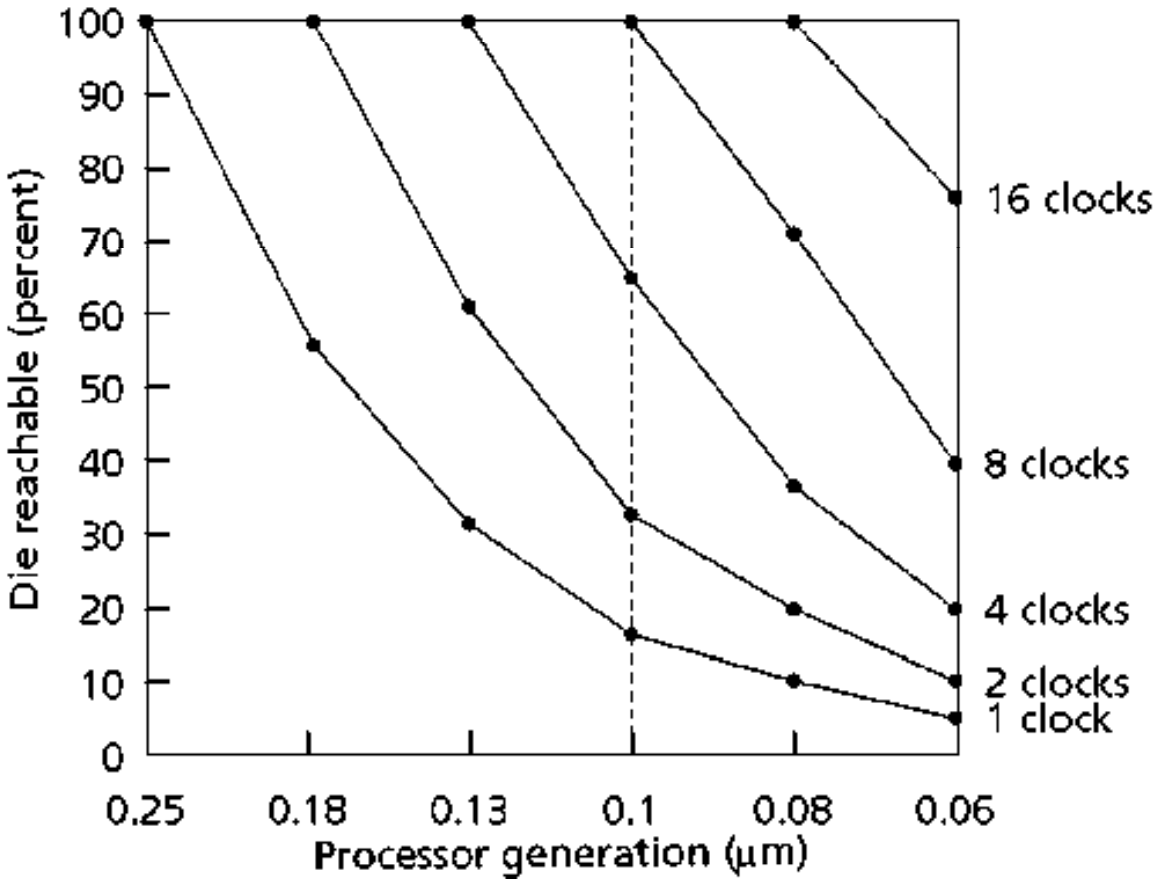
Node	0.6 um	0.06 um	ratio
Die length	16 mm	32 mm	2 X
Gates on Die	1 million	400 million	400 X
Clock Frequency	166 MHz	2.5 GHz	15 X
t_{gate}	250 ps	15.6 ps	16 X
$d_{wire\ locality(raw)}$	5 mm	0.03125 mm	1/160 X
$d_{wire\ locality(improv)}$	5 mm	0.125 mm	1/40 X
$d_{clock\ locality}$	Die if >.18 µm	1.5 mm	1/40 X
Gates in Region	100,000	6,000	1/16 X

Assumptions over 8 process steps:

- 150% increase in gate speed per process step
- 20% wire improvement per process step
- 10% die size increase per process step

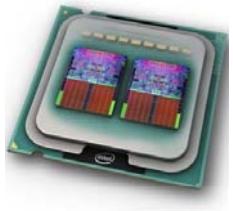
Die reachable per clock





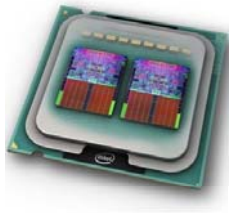
Process nodes: .6, .35, .25, .18, .13, .1, .08 .06 µm

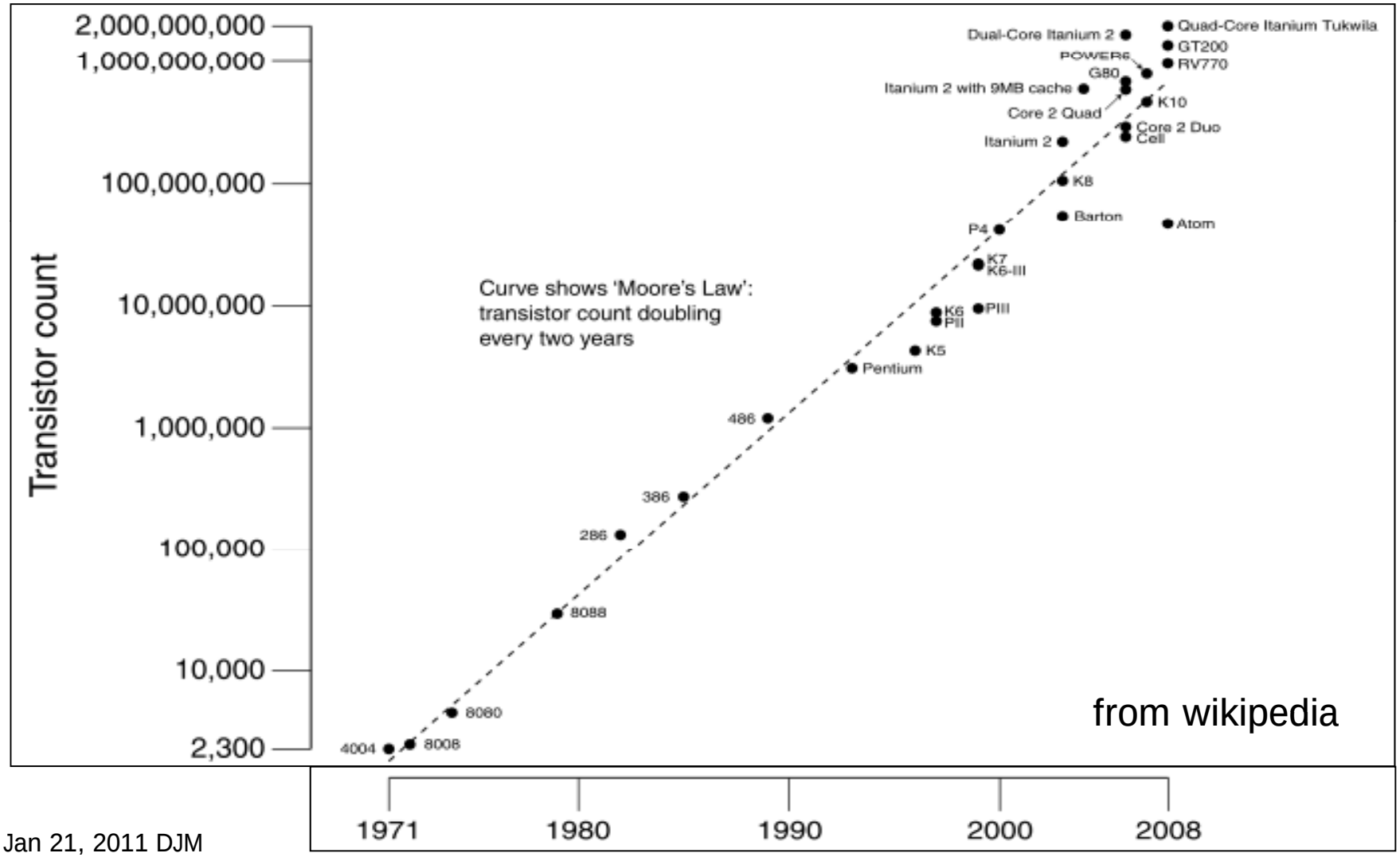
Trends Since 1997 paper



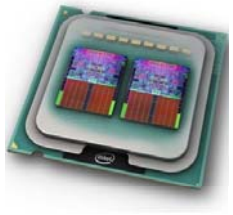
- Clock speeds have maxed out ~3 GHz
- Moore's law w/high dielectric materials
- Process nodes are now at 32 nm (next 22)
- 10 chips since 2003 w/ > 1 B transistors
- Multiple CPU chips are the norm
- Large fine grain GPU and FPGA chips
- Power major design constraint (>200 W)

Transistor Counts 1971-2008



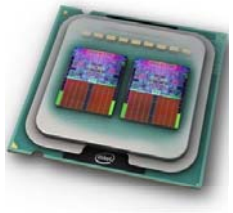


Billion Transistor Chips by 2010

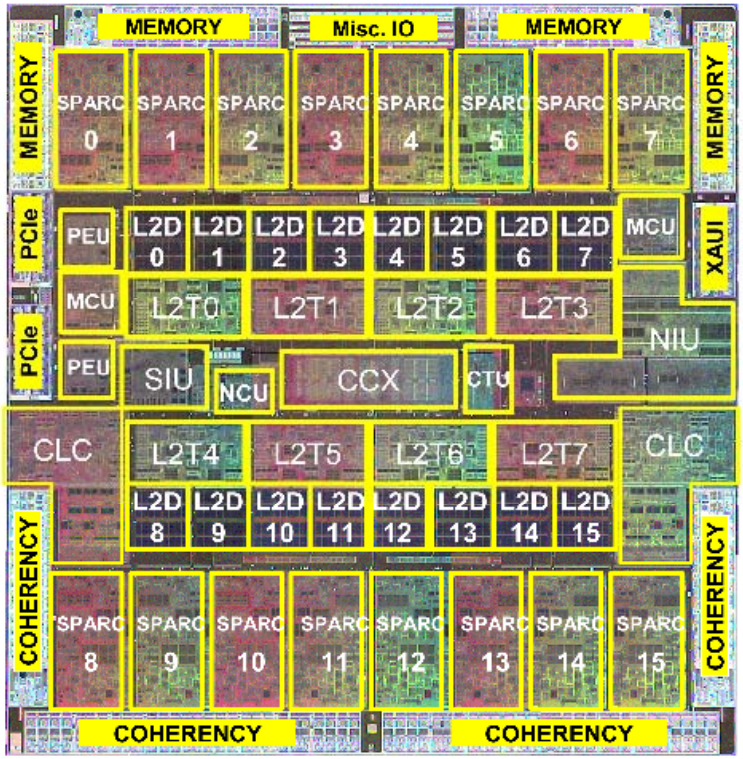


Product	Date	Trans	Proc	Cores	Codename	Developer
Itanium	Feb 2011	3.1 B	32 nm	8*4	Poulson	Intel
Nvidia GTX 570	Dec 2010	3 B	40 nm	480	GF110	NVIDIA/GPU
UltraSPARC T3	Sep 2010	1 B	45 nm	16*8	Niagara 3	Sun/Oracle
Core i7-980X	Mar 2010	1.17 B	32 nm	6*2	Gulftown	Intel
Intel Xeon	May 2009	2.3 B	45 nm	8*2	Beckton	Intel
Intel Itanium	Feb 2008	2 B	65 nm	4	Tukwila	Intel
Power7 (8-core)	Aug 2009	1.2 B	45 nm	8*4	Power 7	IBM
GeForce GTX 280	Dec 2008	1.4 B	65 nm	240	GPX 200	NVIDIA/GPU
Itanium-2	Oct 2005	1.72 B	90 nm	2*2	Montecito	Intel
Stratix IV FPGA	May 2008	2.5 B	40 nm	680K	FPGA Gates	Altera
Virtex FGPA	Sep 2003	1 B	70 nm	4 PPC	FPGA w/PPC	Xilinx

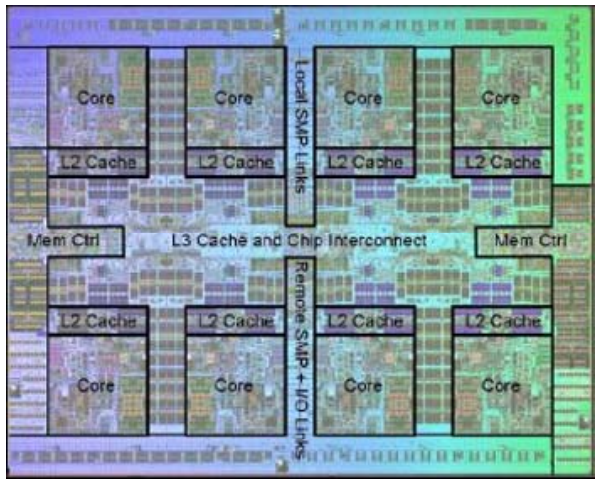
Core Wars



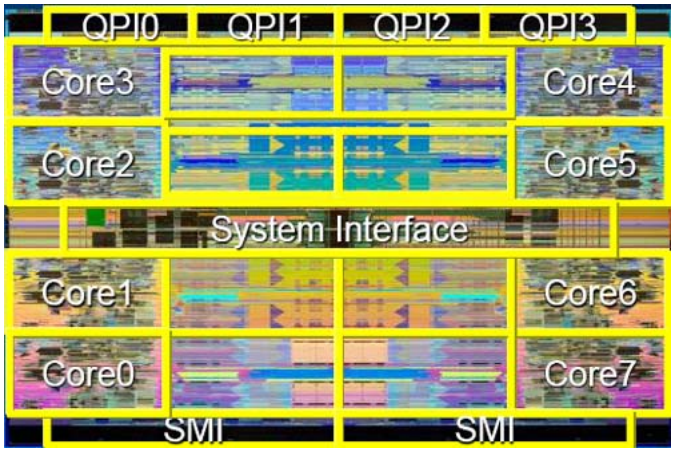
Sun Niagara 3: 16 Cores



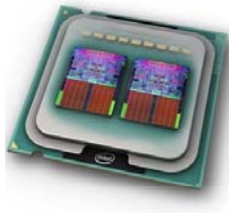
IBM_Power7: 8 cores



Intel Xeon “Beckton”: 8 Cores



ITRS: International Technology Roadmap for Semiconductors



15 year forecast from 2003 ITRS - International Technology Roadmap for Semiconductors at: http://www.itrs.net/

These sizes are close to physical limits and technological limits.

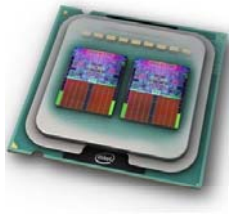
Near-term Years

YEAR OF PRODUCTION	2003	2004	2005	2006	2007	2008	2009
Technology Node		hp90			hp65		
DRAM ½ Pitch (nm)	100	90	80	70	65	57	50
MPU/ASIC M1 ½ Pitch (nm)	120	107	95	85	75	67	60
MPU/ASIC Poly Si ½ Pitch (nm)	107	90	80	70	65	57	50
MPU Printed Gate Length (nm)	65	53	45	40	35	32	28
MPU Physical Gate Length (nm)	45	37	32	28	25	22	20

Long-term Years

YEAR OF PRODUCTION	2010	2012	2013	2015	2016	2018
Technology Node	hp45		hp32		hp22	
DRAM ½ Pitch (nm)	45	35	32	25	22	18
MPU/ASIC M1 ½ Pitch (nm)	54	42	38	30	27	21
MPU/ASIC Poly Si ½ Pitch (nm)	45	35	32	25	22	18
MPU Printed Gate Length (nm)	25	20	18	14	13	10
MPU Physical Gate Length (nm)	18	14	13	10	9	7

Updated ITRS Forecast

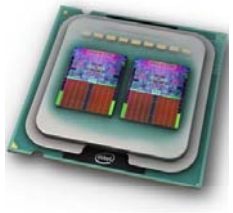


	Year of Production	2009 "32nm"	2010 "22nm"	2012	2013 "16nm"	2015	2016 "11nm"		
2010 ORTC	Flash ½ Pitch (nm) (un-contacted Poly)(f)[A]	38	32	28	25	23	20	18	15.9
2010 PIDS Projection based on survey data	Flash ½ Pitch (nm) (un-contacted Poly)(f) [B]	N/A	26	24	22	20	19	18	16
2010 WAS	DRAM ½ Pitch (nm) (contacted)[C]	52	45	40	36	32	28	25	22.5
2010 PIDS Projection based on survey data	DRAM ½ Pitch (nm) (contacted) [D]	N/A	42	36	31	28	25	24.0	21.0
	MPU/ASIC Metal 1 (M1) ½ Pitch (nm)[1,2]	54	45	38	32	27	24	21	18.9
	MPU Printed Gate Length (GLpr) (nm) ††[1]	47	41	35	31	28	25	22	19.8
	MPU Physical Gate Length (GLph) (nm)[1]	29	27	24	22	20	18	17	15.3

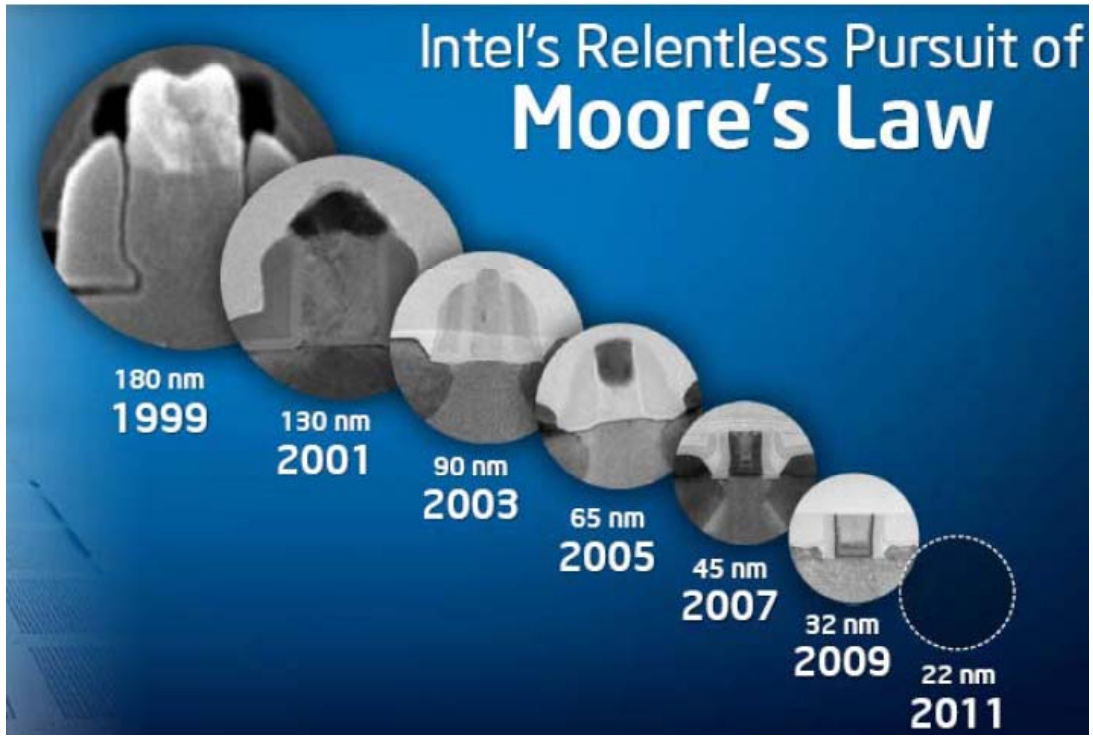
2017	2018	2019	2020	2021	2022	2023	2024	2025
"8nm"								
14.2	12.6	11.3	10.0	8.9	8.0	7.1	6.3	N/A
14	13	12	11	9	8	8	8	N/A
20.0	17.9	15.9	14.2	12.6	11.3	10.0	8.9	N/A
18.0	16.0	14.0	13.0	12.0	10.0	9.0	8.0	N/A
16.9	15.0	13.4	11.9	10.6	9.5	8.4	7.5	N/A
17.7	15.7	14.0	12.5	11.1	9.9	8.8	7.9	N/A
14.0	12.8	11.7	10.7	9.7	8.9	8.1	7.4	N/A

15 year forecast from 2009 ITRS - International Technology Roadmap for Semiconductors-updated: http://www.itrs.net/

Intel's Process Roadmap



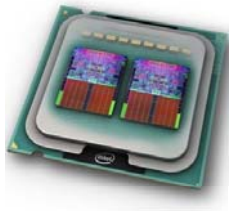
90nm	65nm	45nm	32nm	22nm	15nm	11nm	08 ?
2003	2005	2007	2009	2011	2013	2015	2017



Source:
Paul Otellini, Intel CEO, "Building a Continuum of Computing", Opening Keynote, Intel Developer Forum 2009, San Francisco, Sep 22 – 24, 2009

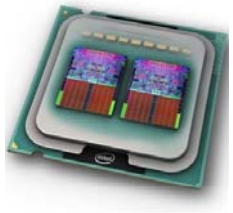
See HANS STORK IEEE NanoTech 2010 paper

Computer Scaling Limits



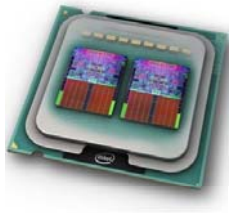
- Physical Limits
  - Power density/Dissipation: max is 100 W/cm^2
  - Thermal/noise: E/f = 100h
  - Molecular/atomic/charge discreteness limits
  - Quantum: tunneling & Heisenberg uncertainty
- Technology Limits
  - Gate Length: min > 8 nm (with new materials)
  - Lithography Limits: wavelength of visible light
  - Power dissipation (<100 watts) and Temperature
  - Wire Scaling: multicpu chips at ~ billion transistors
  - Materials for dielectrics etc

2010 Design Trends

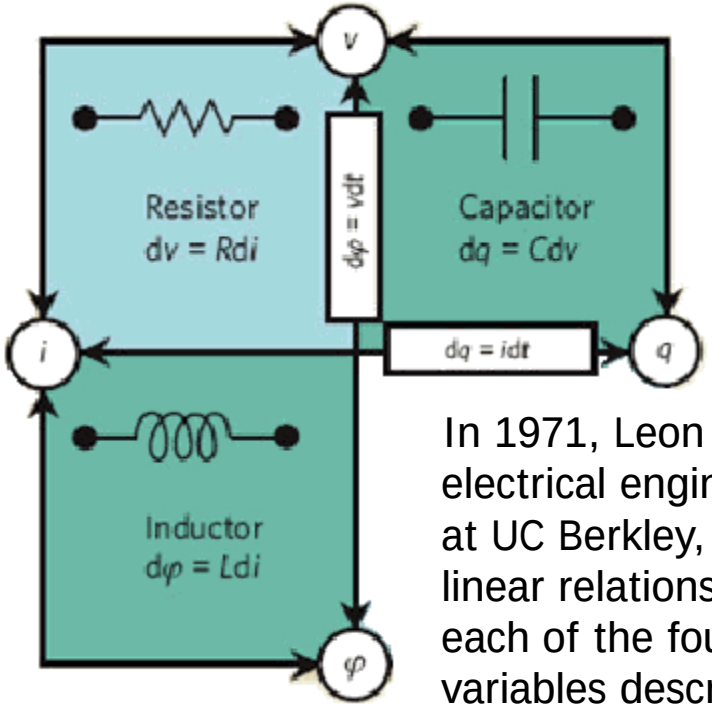


- Multicpu Chips will continue
  - Manage power & no more clock increases
  - Requires innovation in parallel computing
  - Designs may top at < 100 cpus
  - CPUs and GPUS integrated (Sandy Bridge at Intel)
- Higher density/lower power solutions
  - DSP/CPUs heterogeneous systems for portable systems
  - CPU/FPGA systems (Convey Computers, IBM)
  - New memory/logic devices (spintronics)
  - Memristor based systems (HP, Numenta)
- Quantum Computing is nitch market

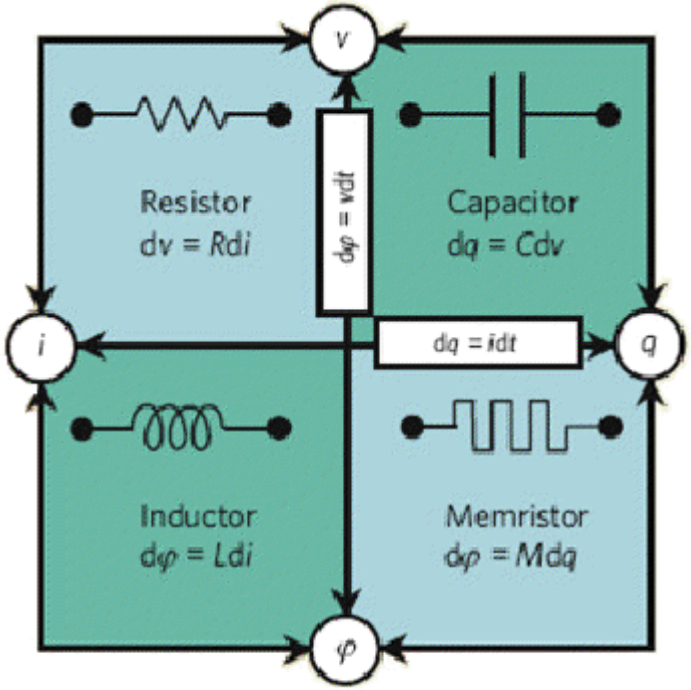
Memristor Fundamentals



Three original 2-terminal circuit elements (based on current, voltage, charge, and magnetic flux relationships)

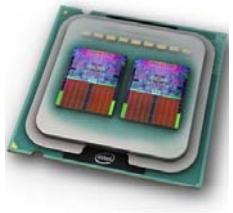


In 1971, Leon Chua, an electrical engineer professor at UC Berkley, arranged the linear relationships between each of the four basic variables describing the above circuit relationships.



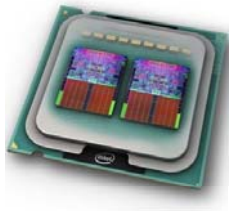
four final 2-terminal circuit elements

Scaling Predictions



- Semiconductors will stop scaling in <10 yrs
  - Nanocomputers won't stop this; only delay it
  - Breakthroughs required or industry stagnates
  - College students consider non-semiconductor careers
- High dimensional Research in other areas:
  - Deep meaning and automatic learning
  - Programming probabilistic parallel computers
  - Noise as valued resource instead of unwanted
  - Higher dimensional computing
  - Investigate non-local computing
  - Biological inspired computing – Quantum Brain?

Summary



- Predictions in ‘97 came true as expected
- Scaling wall is now visible to industry
- Heat limits my stop multiprocessor count
- Materials innovation allows more of Moore
- New devices may help scaling (more than Moore)
- Fab Costs may slow before physical limits
- Must think outside 3d box (quantum?)
- Watch for unexpected aspects of qunoise
- Tablet/phone computing changes markets
- Clouding computing virtualization trends

Questions and Discussions