SAND2010-6124C

Photonic Interconnection Technology and Computer Architecture Scaling

Erik P. DeBenedictis

with section by Scott Hemmert

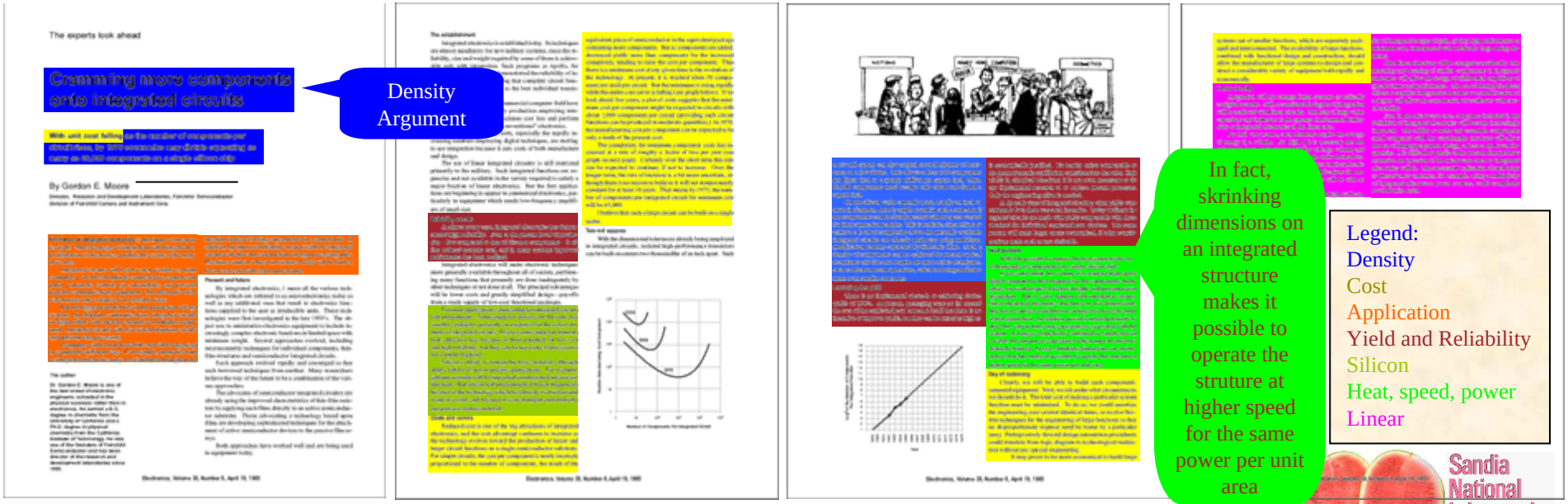
Sandia National Laboratories

Key Points

- Expectations are set by “Moore’s Law” scaling, but it shifts due to marketing pressures
- There was a shift in emphasis from high speed to low power, but the low-power trend will continue to the extreme condition of no power.
- For enabling technologies, including photonics:
  - Logic scaling must include decreasing duty cycle over time
  - Important optical interconnect applications must scale to lower power with decreasing duty cycle

The New Moore's Law

- **The following text appears in the original article and highlights the role of marketing in redfining Moore's Law:**
  - **"In fact, skrinking dimensions on an integrated structure makes it possible to operate the struture at higher speed for the same power per unit area"**
- **How many scientists in the room could get better research results if they were able to redefine e. g. the speed of ligh**



Moore's Paper from Electronics magazine 1965 with topics highlighted

Implications of Moore's Law to be Discussed

- Moore's paper from 1965 implied a basic knowledge of CMOS scaling
  - Later to be →
- Industry expectations were set by Moore's Law
- However, VDD not scaling as expected
- Capicatance model no longer holds
- Software has bursty behavior

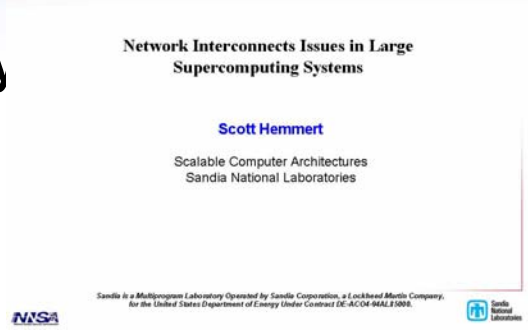
- Dennard Scaling
  - Area $1/\kappa^2$
  - Capacitance $1/\kappa$
  - Resistance κ
  - Threshold (V_{th}) $1/\kappa$
  - Current (I_d) $1/\kappa$
  - Gate Delay (τ_{gd}) $1/\kappa$
  - Wire Delay (τ_{wire}) 1
  - Power $1/\kappa^2 \rightarrow 1/\kappa^3$

Agenda

- **Interconnect matters – study using Red Storm**
- **Voltage scaling**
- **Layer scaling**
- **Faster computing with the power turned off**
- **CMOS interconnect limits**
- **Conclusions for optical interconnect**

System-level Interconnect and Energy

Source:


Network Interconnects Issues in Large Supercomputing Systems

Scott Hemmert

Scalable Computer Architectures
Sandia National Laboratories

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.

- System-level interconnect performance is the key determining how well many applications scale
- With increasing bandwidths, interconnect power real concern
  - Serdes don't turn off well (OK, they turn off fine, they just don't turn back on quickly, due to channel initialization times)
    - Uses power whether valid data is moving through the network or not
- A lot of discussion lately on minimizing picojoules/bit
- However, interconnects are not used in isolation and a system view is vital to maximizing energy efficiency
  - NIC and router architectures, topologies and MPI implementations all play an important role

The Exascale Power Challenge

- DOE would like an exascale machine for 20 MW
  - That's a mere $20 million/year for electricity
- Floating-point power
  - Exascale hardware report: ~6-7 pJ/FLOP
  - So, 1 EF/s = 6-7 MW, just for the floating point operation
    - Does not include rest of processor core or data movement
- Memory power
  - Assumptions: ½ byte/FLOP memory bandwidth, 10 pJ/bit memory access
  - 500 PB/s memory bandwidth = 40 MW
- Network power
  - Assumptions: ¼ byte/FLOP network bandwidth, 2 pJ/bit transmission power, ¼ of switch BW to hosts
  - 625 PB/s total network bandwidth = 10 MW (transmit/receive only)

Application Case Study: CTH

- **CTH is a multi-material, large deformation, strong shock wave, solid mechanics code developed at Sandia National Laboratories. CTH has models for multi-phase, elastic viscoplastic, porous and explosive materials.**

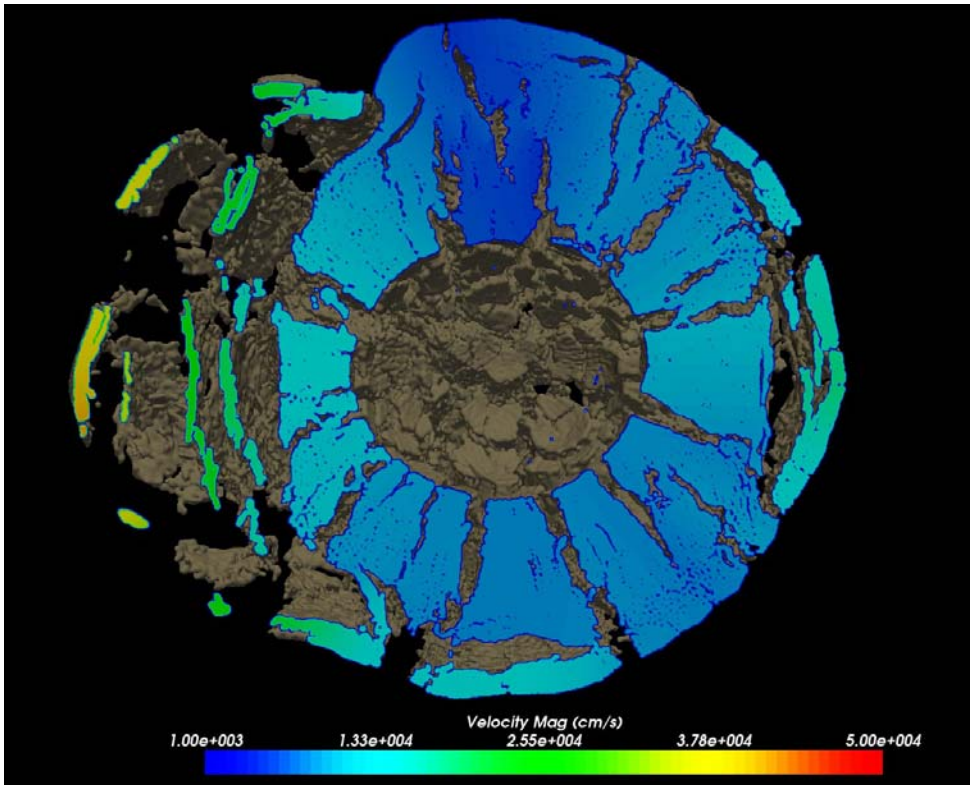
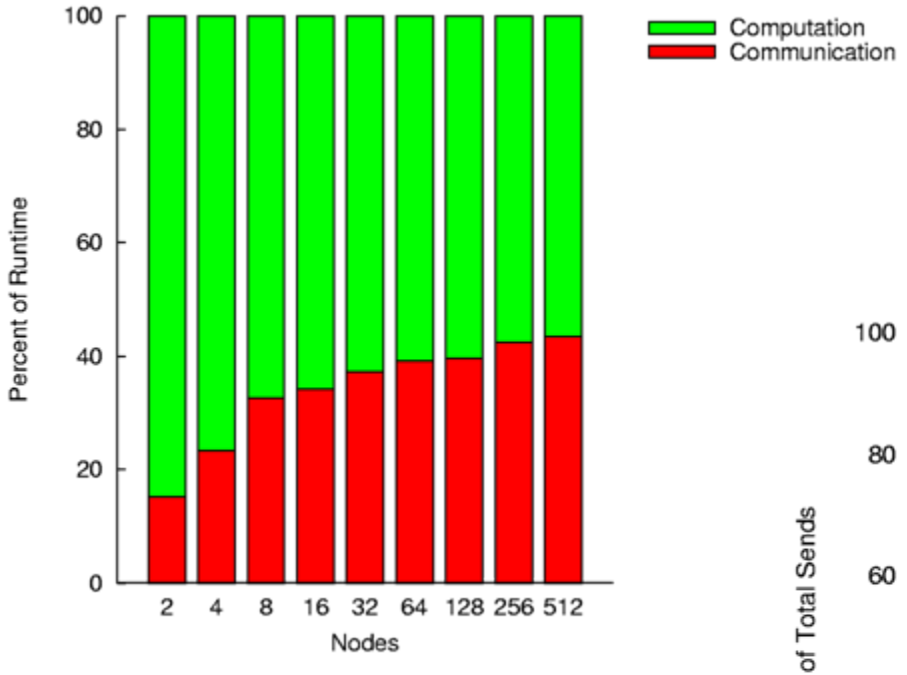


Image courtesy of ASC

Asteroid Golevka measures about 500 x 600 x 700 meters. In this CTH shock physics simulation, a 10 Megaton explosion was initiated at the center of mass. The simulation ran for about 15 hours on 7200 nodes of Red Storm and provided approximately 0.65 second of simulated time.

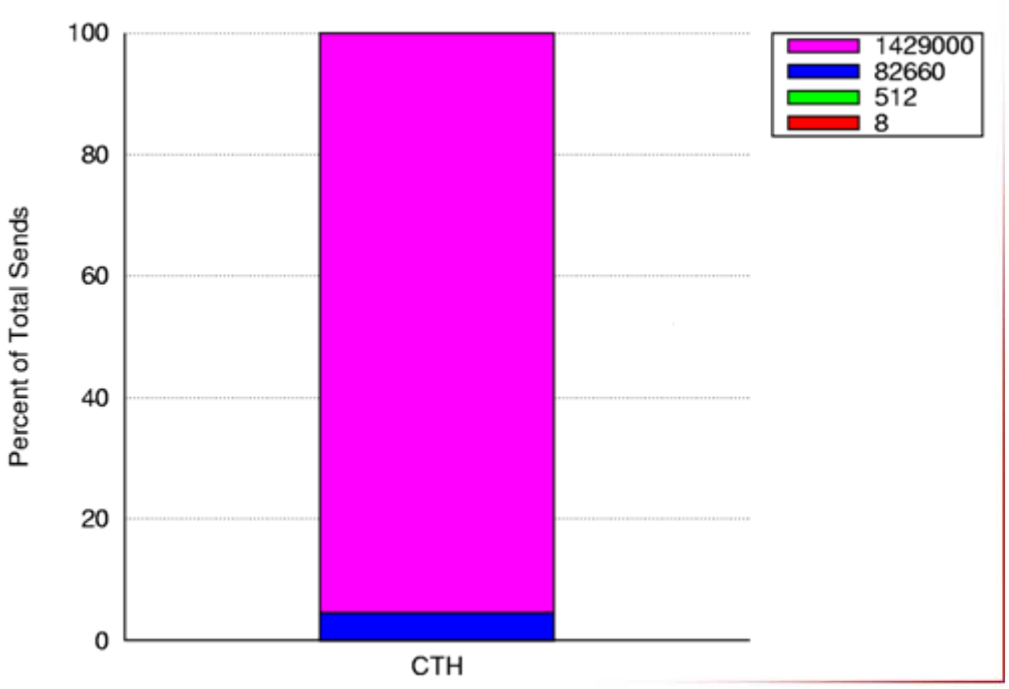
Application Case Study: CTH

Shaped Charge Problem (weak scaling)



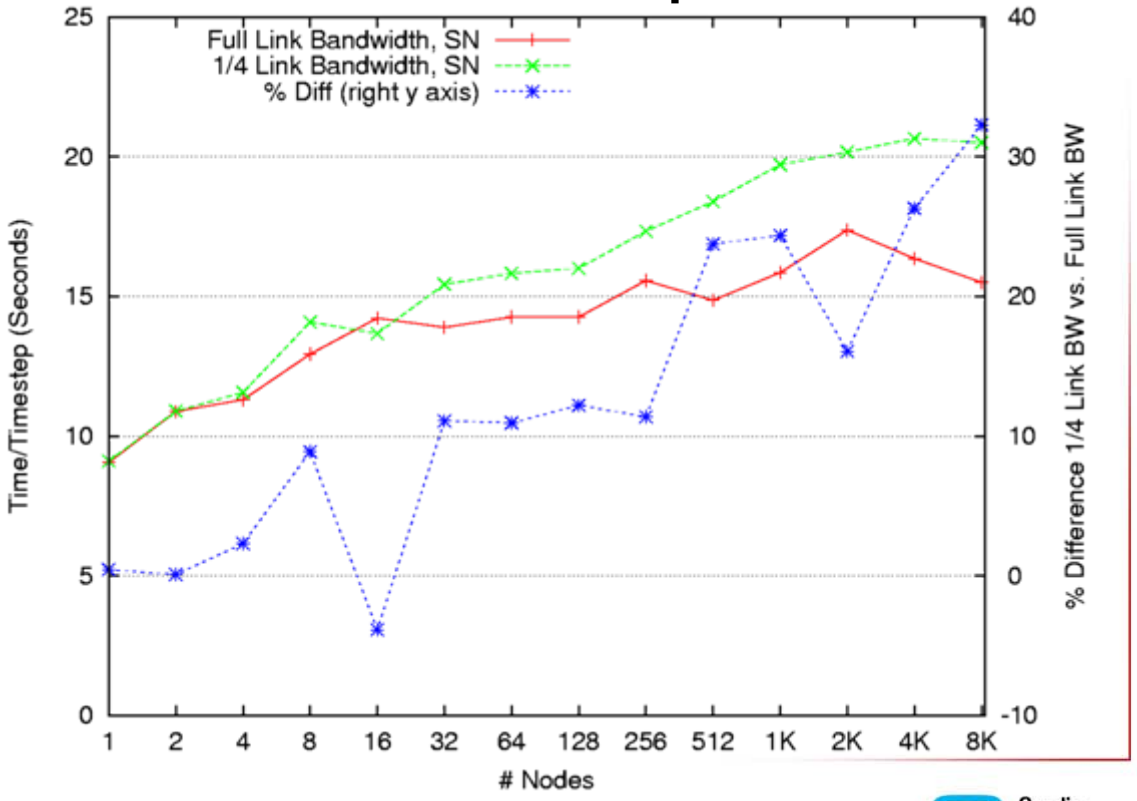
As job size increases, communication time can grow to consume around 40-50% of the runtime.

CTH communication is dominated by long messages.



CTH Bandwidth Degradation Study

- **Uses capabilities built into the Red Storm SeaStar interconnect to turn off interconnect router lanes at boot time**
  - **Links are made up of 4 3-bit subchannels that can be independently enabled**
- **Measure application performance at full and one-quarter link bandwidth**
- **At largest measured job size, quartering bandwidth leads to 32% longer runtime**

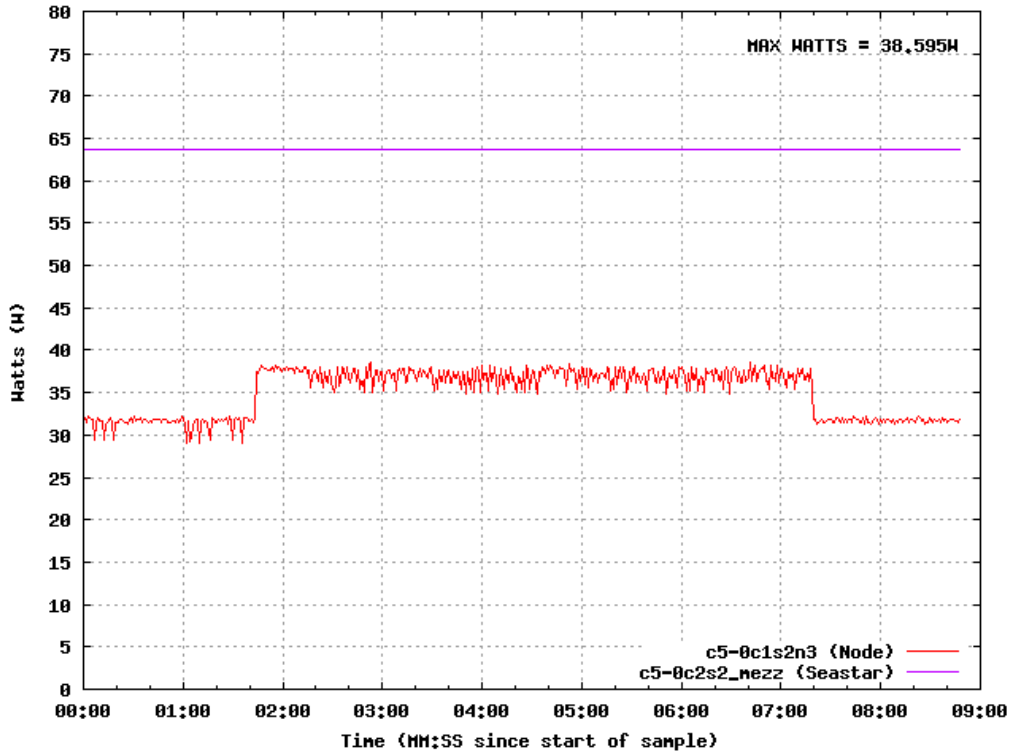


CTH Power Signature Study

- **Power measured using Red Storm's built-in current monitors**

Total Node Power:

CPU: 37 (red)
SeaStar: 16 (blue/4)
Memory: 20 (estimated)
73 Watts



Putting it all Together

- **Assume interconnect power drops linearly with bandwidth**
  - **68% of the performance for 25% of the interconnect power**
- **Total power for ¼ bandwidth = 61 Watts (down from 73 watts)**
  - **68% of the performance for 83.6% of the system power**

- **Total Energy for two cases assuming full bandwidth runtime of X**

$$\text{Energy}_{\text{full}} = 73X \qquad \text{Energy}_{1/4} = 1.32X * 61 = 80.5X$$

$$\frac{\text{Energy}_{1/4}}{\text{Energy}_{\text{full}}} = \frac{80.5X}{73X} = 1.10$$

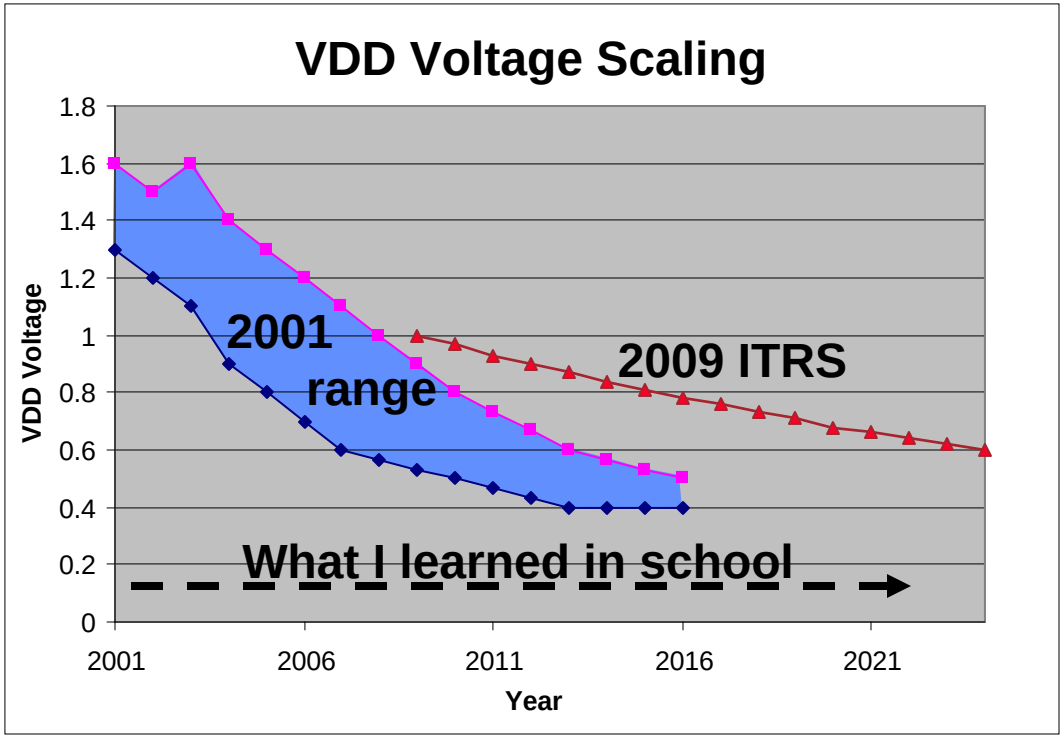
- **Net energy increase of 10% for ¼ bandwidth case**
  - **Keep in mind this doesn't count the energy used for the file system attached to the machine or other machine room costs**

Agenda

- Interconnect matters – study using Red Storm
- Voltage scaling
- Layer scaling
- Faster computing with the power turned off
- CMOS interconnect limits
- Conclusions for optical interconnect

ITRS Voltage Scaling

- When I went to school in the 1970s, VDD was project to expected to drop to ~130mv
- ITRS was predicting .4-.5 volts as of 2001
- Current predictions are much higher
- Energy is $\frac{1}{2}CV^2$



Sandia National Laboratories

Data (Backup)

ITRS CMOS scaling is the result of device simulations using a program called MASTAR coupled with computer system models. Everything evolves.

YEAR OF PRODUCTION	2001	2002	2003	2004	2005	2006	2007	2010	2013	2016
DRAM ½ PITCH (nm)	130	115	100	90	80	70	65	45	32	22
MPU / ASIC ½ PITCH (nm)	150	130	107	90	80	70	65	50	35	25
MPU PRINTED GATE LENGTH (nm)	90	75	65	53	45	40	35	25	18	13
MPU PHYSICAL GATE LENGTH (nm)	65	53	45	37	32	28	25	18	13	9
Physical gate length high-performance (HP) (nm) [1]	65	53	45	37	32	28	25	18	13	9
Equivalent physical oxide thickness for high-performance T_{ox} (EOT)(nm) [2]	1.3–1.6	1.2–1.5	1.1–1.6	0.9–1.4	0.8–1.3	0.7–1.2	0.6–1.1	0.5-0.8	0.4-0.6	0.4-0.5
Gate depletion and quantum effects electrical thickness adjustment factor (nm) [3]	0.8	0.8	0.8	0.8	0.8	0.8	0.5	0.5	0.5	0.5
T_{ox} electrical equivalent (nm) [4]	2.3	2.1	2.0	2.0	1.9	1.9	1.4	1.2	1.0	0.9
Nominal power supply voltage (V_{dd}) (V) [5]	1.2	1.1	1.0	1.0	0.9	0.9	0.7	0.6	0.5	0.4

Table ORTC-6 Power Supply and Power Dissipation

Year of Production	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Flash ½ Pitch (nm) (un-contacted Poly) (f)	38	32	28	25	23	20	18	15.9	14.2	12.6	11.3	10.0	8.9	8.0	7.1	6.3
DRAM ½ Pitch (nm) (contacted)	52	45	40	36	32	28	25	22.5	20.0	17.9	15.9	14.2	12.6	11.3	10.0	8.9
MPU/ASIC Metal 1 (M1) ½ Pitch (nm)	54	45	38	32	27	24	21	18.9	16.9	15.0	13.4	11.9	10.6	9.5	8.4	7.5
MPU Printed Gate Length (GLpr) (nm) ††	47	41	35	31	28	25	22	19.8	17.7	15.7	14.0	12.5	11.1	9.9	8.8	7.9
MPU Physical Gate Length (GLph) (nm)	29	27	24	22	20	18	17	15.3	14.0	12.8	11.7	10.7	9.7	8.9	8.1	7.4
Power Supply Voltage (V)																
Vdd (high-performance)	1.0	0.97	0.93	0.9	0.87	0.84	0.81	0.78	0.76	0.73	0.71	0.68	0.66	0.64	0.62	0.6
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Allowable Maximum Power [1]																
High-performance with heatsink (W)	143	146	161	158	149	152	143	130	130	136	133	130	130	130	Intentionally Blank	Intentionally Blank
Maximum Affordable Chip Size Target for High-performance MPU Maximum Power Calculation [2]	260	260	260	260	260	260	260	260	260	260	260	260	260	260	260	260
Maximum High-performance MPU Maximum Power Density for Maximum Power Calculation	0.46	0.47	0.52	0.51	0.48	0.49	0.46	0.42	0.42	0.44	0.43	0.42	0.42	0.42	0.42	0.42

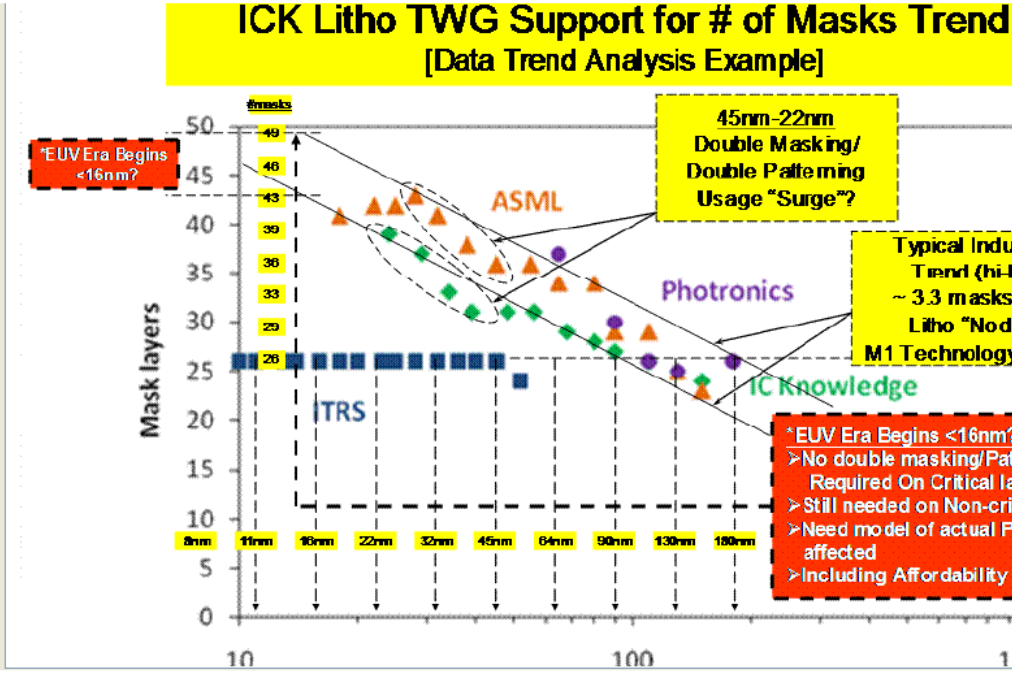
Sandia National Laboratories

Agenda

- Interconnect matters – study using Red Storm
- Voltage scaling
- Layer scaling
- Faster computing with the power turned off
- CMOS interconnect limits
- Conclusions for optical interconnect

Evolving Moore's Law: More Layers

- **Number of masks is now expected to increase, including number of interconnect layers**
- **Per Moore's Law, power per unit area is constant – but that applies to one metal layer**
- **Another pressure to raise power**



• Preliminary feedback:

•No differentiation among different products

•No mask cost increase by node

•Write time increase

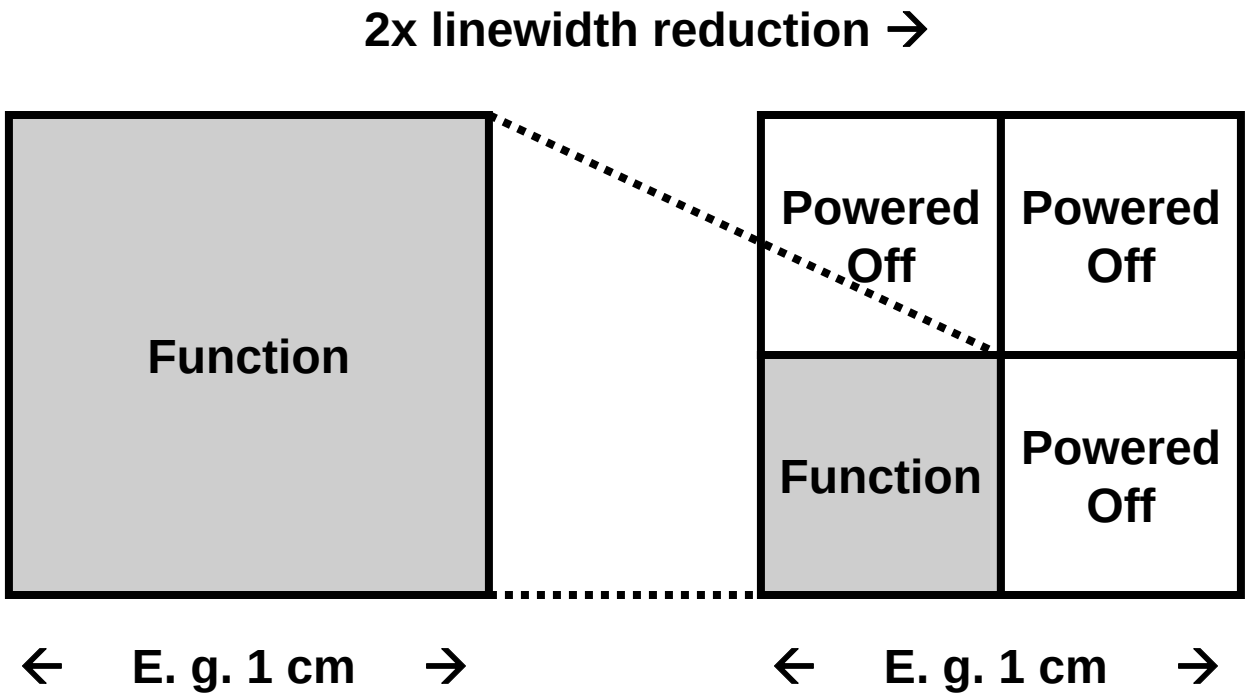
•No breakdown on "non-critical"

• Need to check pu TSMC and Intel

ITRS International Technology Roadmap for Semico

National Laboratories

Main Idea: Think About Power-Off State

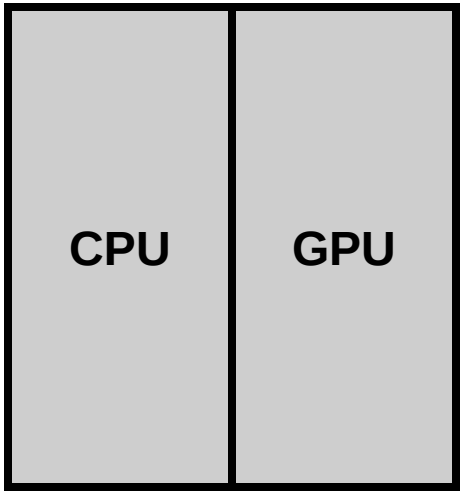


So Moore's Law gives you an additional 3x additional transistors, the complicating factor is that they must be powered off!

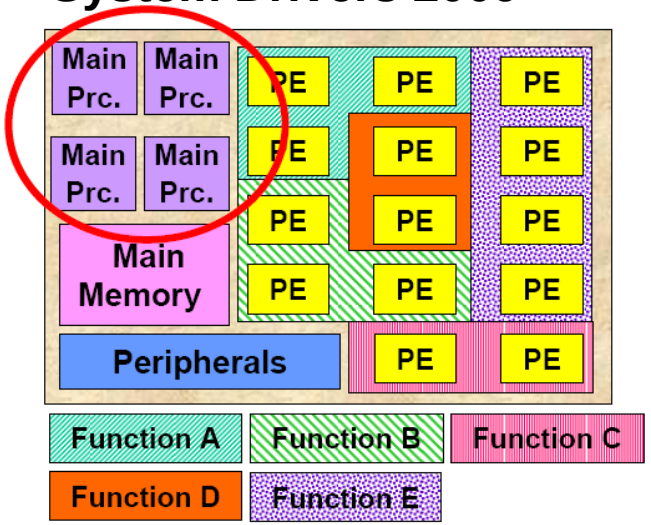
Sandia National Laboratories

How Do You Compute Faster with Powered-off Devices?

Example: CPU/GPU



From ITRS Design and System Drivers 2009

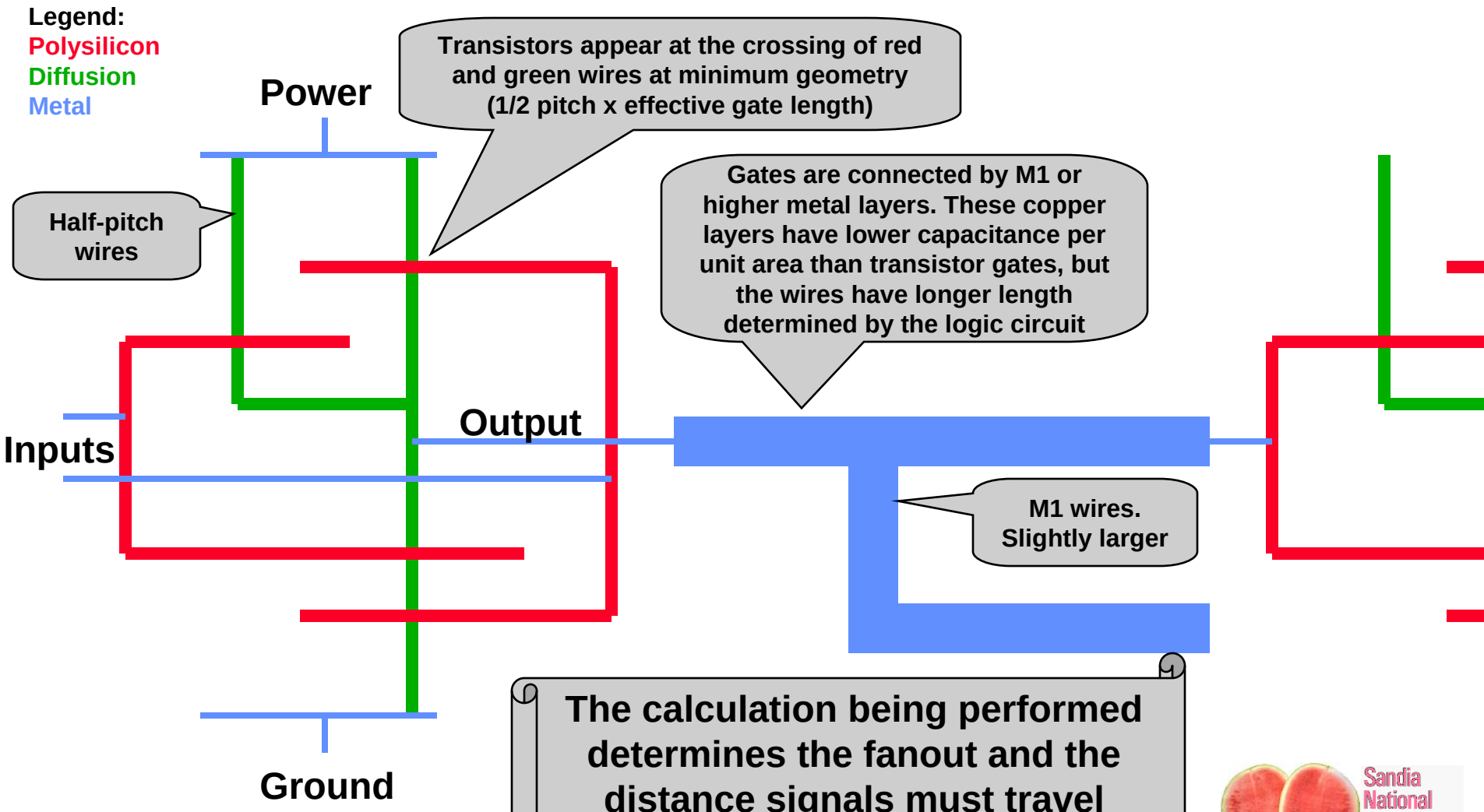


The architectures illustrated tend to be specialized to compute a limited set of functions with high power efficiency.
The ability to dynamically power up and use the most efficient architecture for a task yields a boost in overall power efficiency

Agenda

- Interconnect matters – study using Red Storm
- Voltage scaling
- Layer scaling
- Faster computing with the power turned off
- CMOS interconnect limits
- Conclusions for optical interconnect

CMOS Model

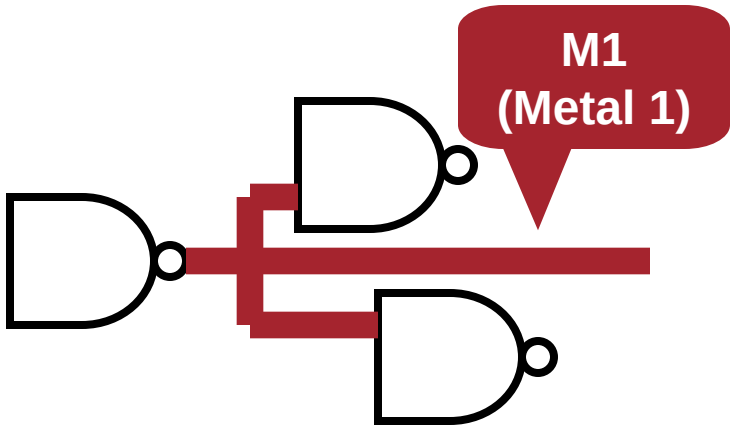


Sandia National Laboratories

Competitive Positioning: ITRS CMOS (the improvement is in the interconnect)

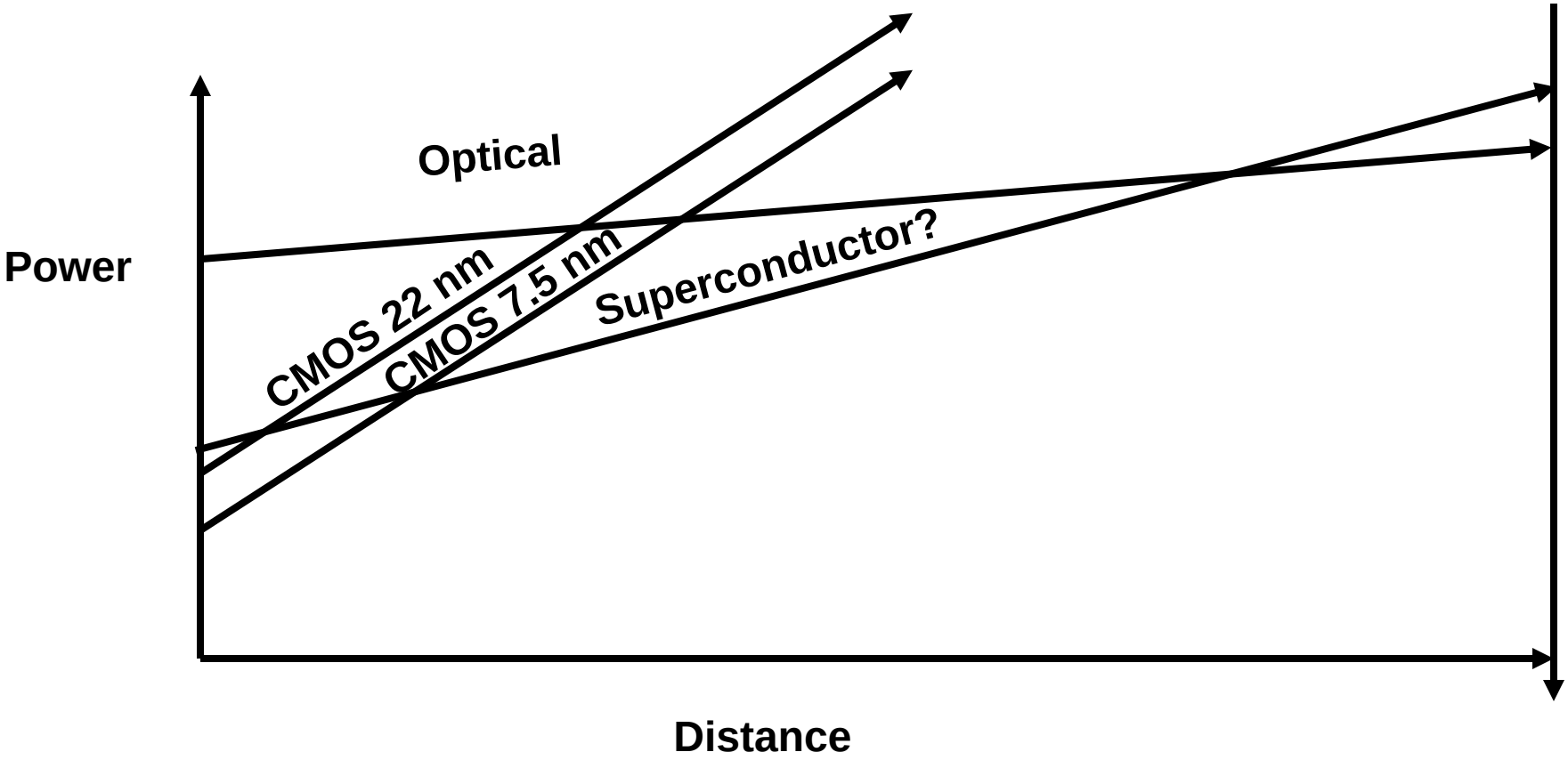
	CMOS 22 nm (CV² for minimum transistor)	CMOS 7.5 nm (CV² for minimum transistor)
Device energy	2500 kT	300 kT
Devices/gate (FO 3)	6	6
Wiring overhead factor	4+	4+
Energy supply overhead	2 (switching power supply losses)	2 (switching power supply losses)
Total	120000+ kT	14400+ kT

For CMOS, wire capacitance is in parallel with gate capacitance. Wire capacitance exceeds gate capacitance by a factor of ~4 even for the most regular arithmetic arrays and by much greater factors otherwise.



Sandia National Laboratories

Interconnect Opportunity Summary



Agenda

- Interconnect matters – study using Red Storm
- Voltage scaling
- Layer scaling
- Faster computing with the power turned off
- CMOS interconnect limits
- Conclusions for optical interconnect

Conclusions: Scaling for Short-Haul Optical Interconnect

- Both static and dynamic power should scale down to match the scaling of the electronics
  - When you reach .5-.05 fJ/bit, talk with me again ☺
- The duty cycle of device operation will scale down as well, raising the importance of low static power
  - Role of heaters?
- Communications will become increasingly bursty, making it important for devices to power up and stabilize quickly
  - PLL's?