Kirk Pruhs

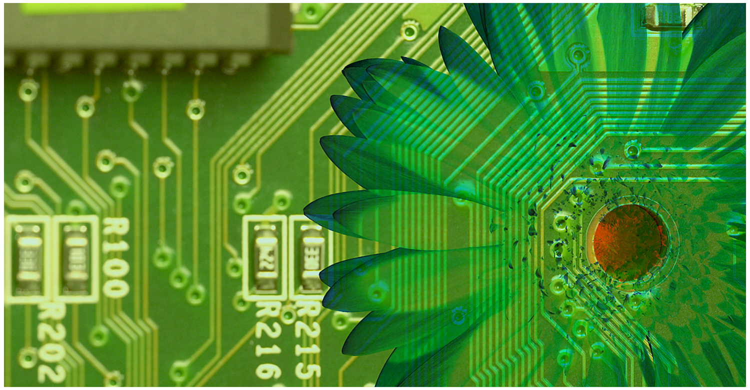
Energy as a Computation Resource

NSF Workshop on Research Directions in the Principles of Parallel Computation

June 28, 2012

Controversial Statement:

- Parallel processing is about more energy efficiency than time efficiency



- "What matters most to the computer designers at Google is not speed, but power, low power, because data centers can consume as much electricity as a city."
  - --- Eric Schmidt, Former CEO Google

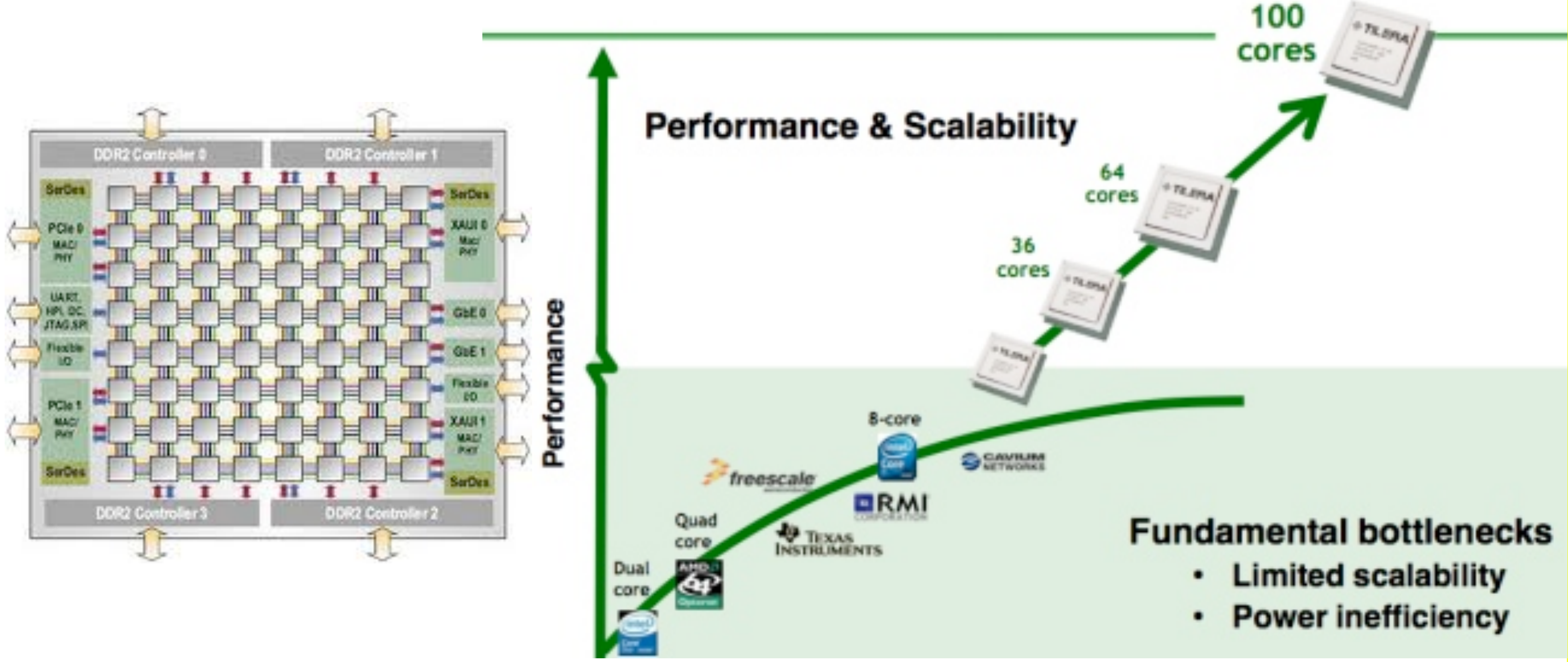
The Masses Now Have Parallel Processing Because of Energy Efficiency Issues

Reuters Friday May 7, 2004

SAN FRANCISCO, May 7 (Reuters) - Intel Corp. said on Friday it has scrapped the development of two new computer chips (code-named Tejas and Jayhawk) for desktop/server systems in order to rush to the marketplace a more efficient chip technology more than a year ahead of schedule. Analysts said the move showed how eager the world's largest chip maker was to cut back on the heat its chips generate. Intel's method of cranking up chip speed was beginning to require expensive and noisy cooling systems for computers.

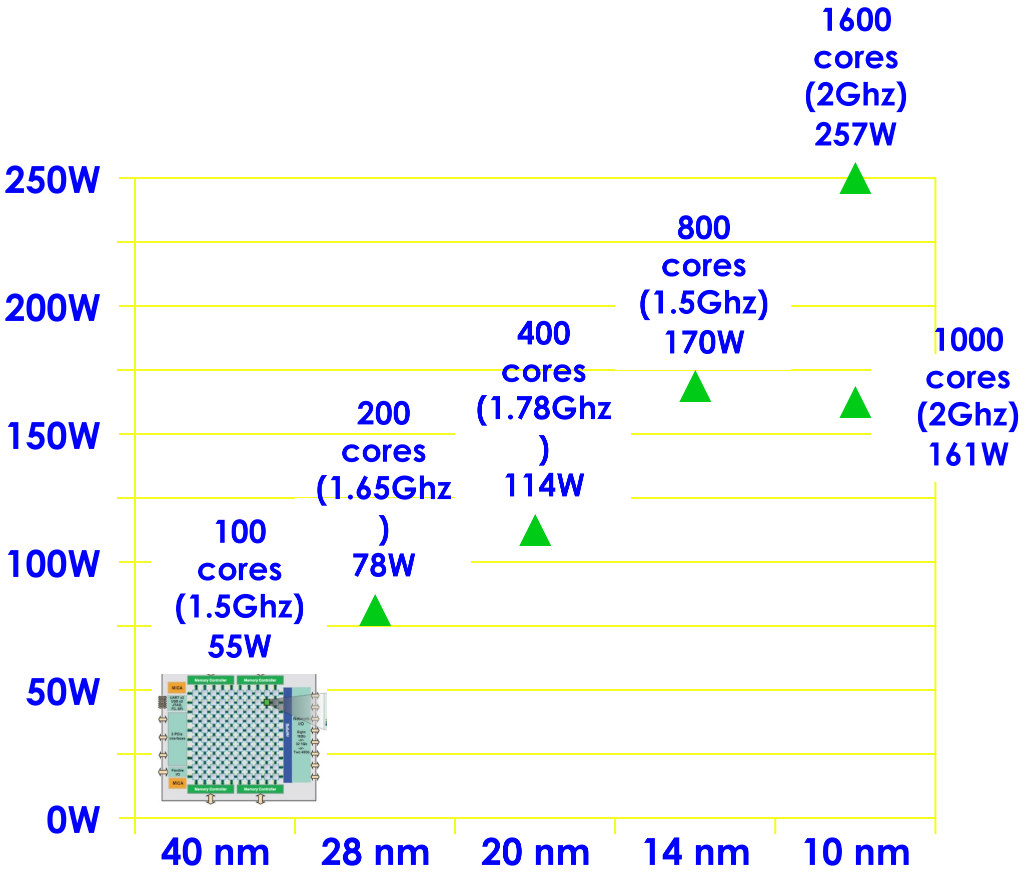
One Vision the Future

- “I would like to call it a corollary of Moore's Law that the number of cores will double every 18 months.”
  - Anant Agarwal, CEO Tilera



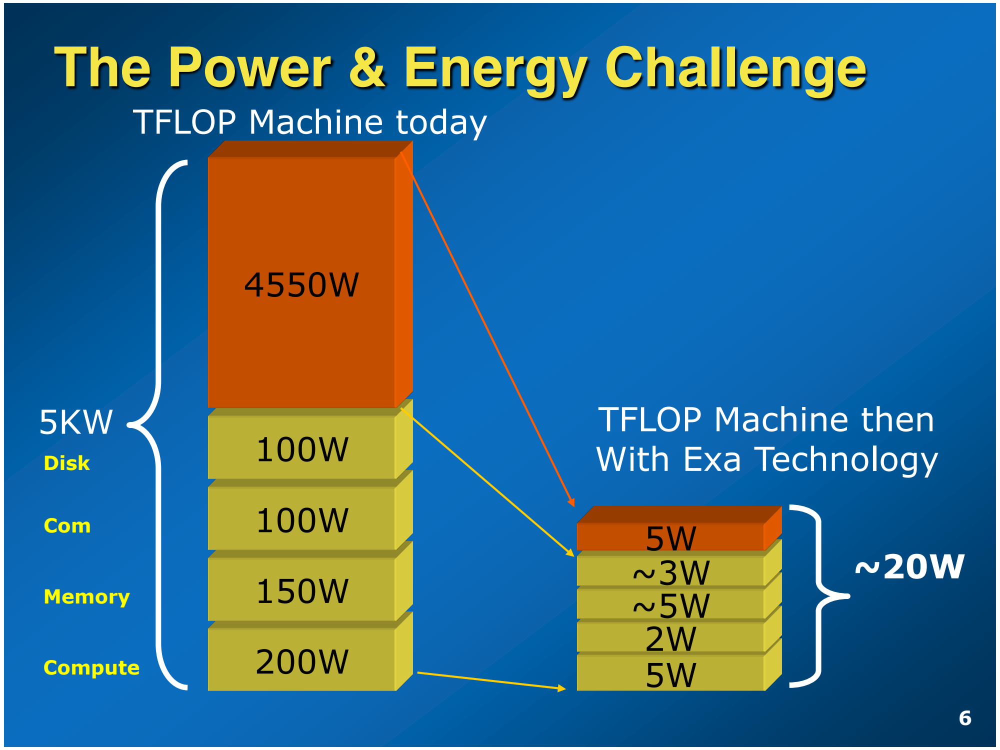
Projections from Bill Lin's keynote at the 2012 Lighter Than Green Dependable Multicore Architectures Workshop

- For thermal reasons, 100W is probably the limit
- The key challenge to get 1000 cores/ processors is energy efficiency



The Exascale Challenge, Shekhar Borkar, Intel

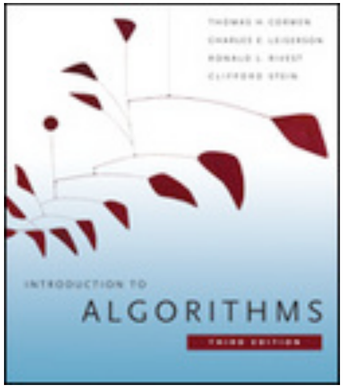
- Key challenge: Two orders of magnitude improvement in energy efficiency

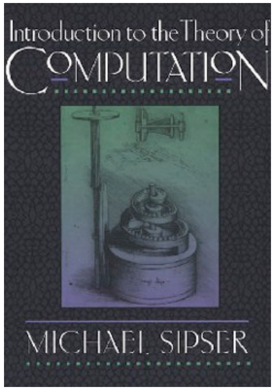


The Power & Energy Challenge

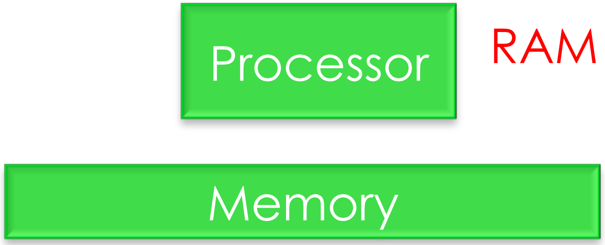
First Question

- What are the right models to abstractly reason about energy, temperature and power as computational resources?



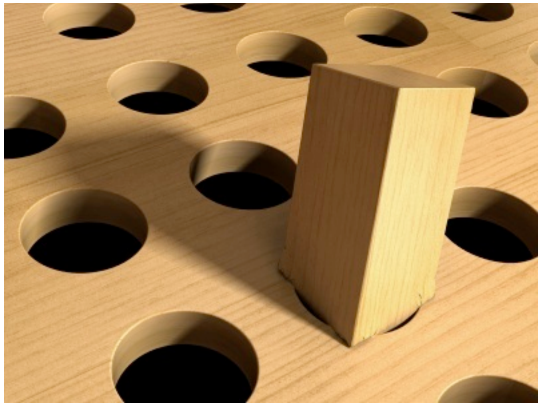


- Reasoning abstractly about time and space as computational resources in simple models has proven to be quite useful



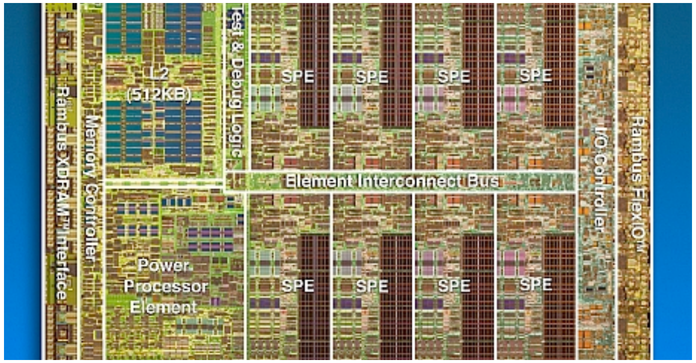
RAM

- However, it seems that due to the fact that the physics of energy is quite different than that of time and space (e.g. there is no minimum energy for computation), we need different models to study energy as a computational resource

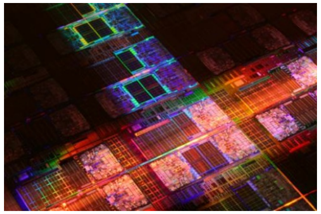


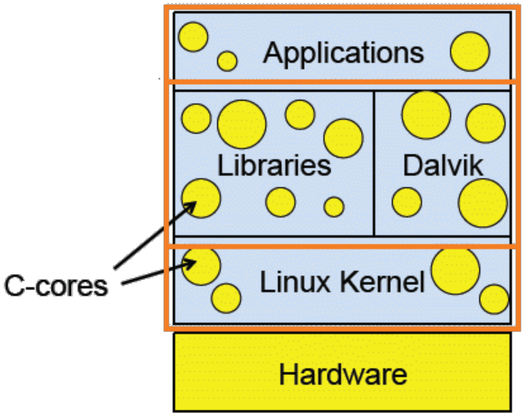
Future Architectures Will Likely have Heterogeneous Processors

- Combination of a few fast processors and many more-energy-efficient slower processors can be an order of magnitude more energy efficient than a homogeneous processor architecture



- Specialized cores that are only powered on when needed may be the best way to combat the "dark silicon apocalypse"





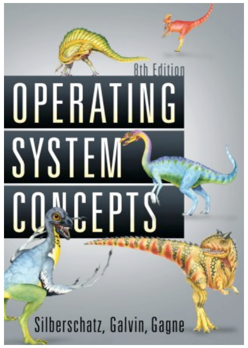
[Source: Goulding, HotChips'10]

Second Question

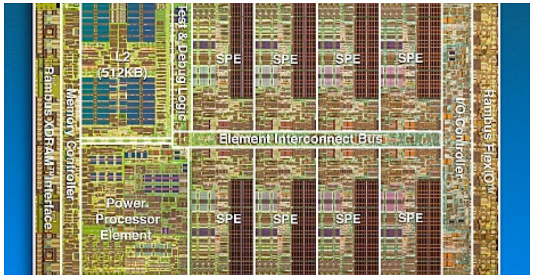
- What are the right models of processor heterogeneity and how does one manage/schedule heterogeneous processors?

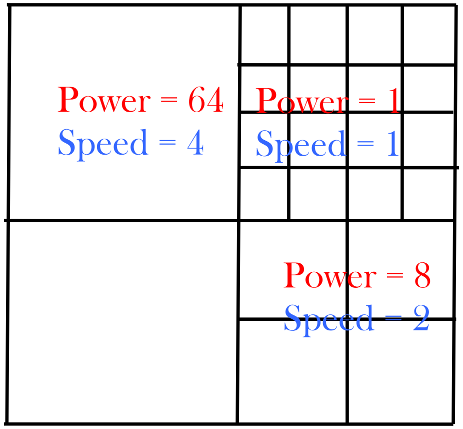
Scheduling On Even Simple Heterogeneous Processors Architectures is Quite Different

- No one knows how to do a worst-case analysis of basically any of the standard priority algorithms, e.g.
  - Shortest Job First
  - Shortest Remaining Processing Time



- Some standard priority algorithms (like Highest Density First) that provably perform well for uniprocessors and homogeneous multiprocessors can perform very badly

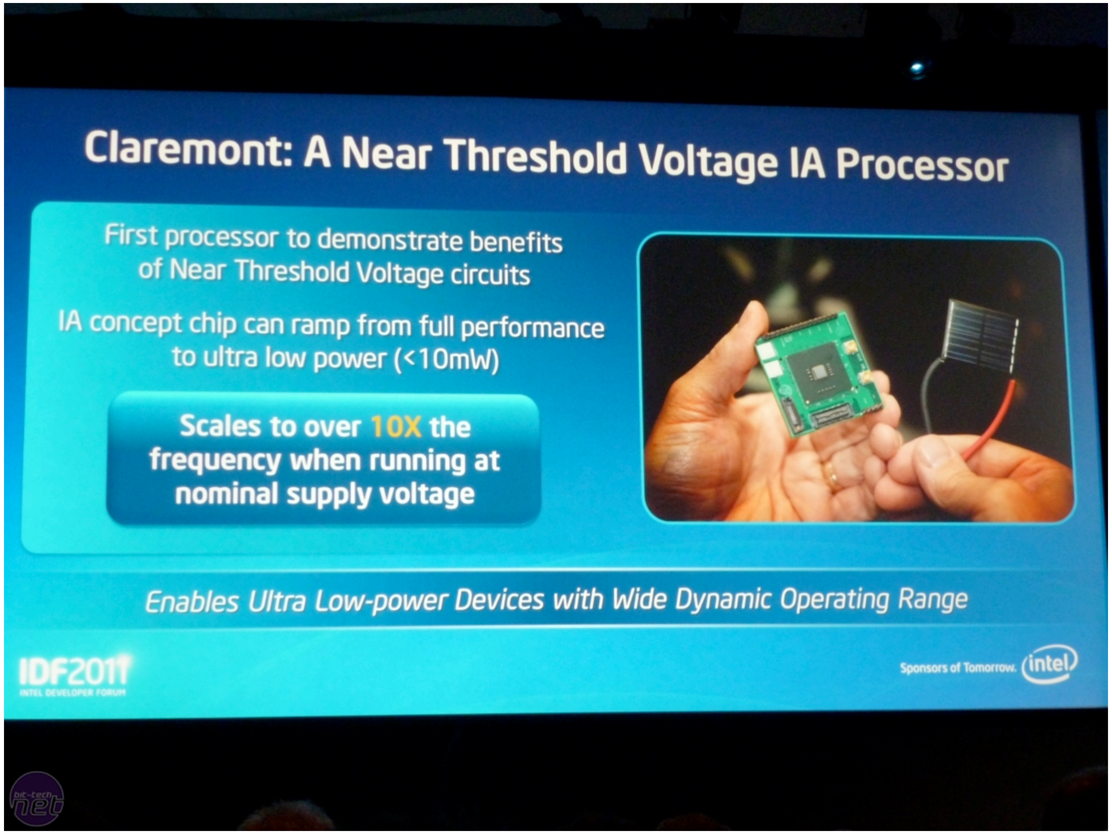




Power = 64
Speed = 4

Power = 1
Speed = 1

Power = 8
Speed = 2



Claremont: A Near Threshold Voltage IA Processor

First processor to demonstrate benefits of Near Threshold Voltage circuits

IA concept chip can ramp from full performance to ultra low power (<10mW)

Scales to over 10X the frequency when running at nominal supply voltage

Enables Ultra Low-power Devices with Wide Dynamic Operating Range

IDF2011 INTEL DEVELOPER FORUM

Sponsors of Tomorrow. intel

- Computing at near threshold voltage increases probability of errors

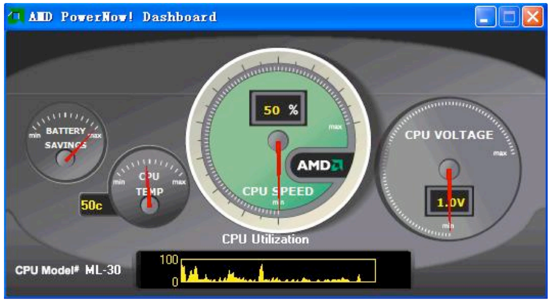
- Also reduced feature size increases probability of errors

Third Question

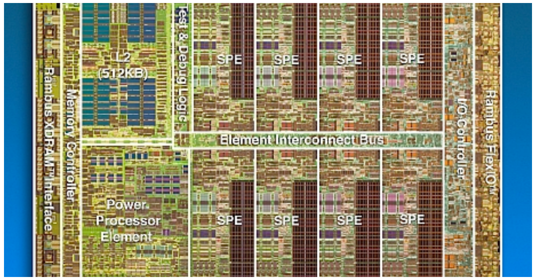
- If computation at the lowest level is more error prone, how should higher levels be designed so that they are more resilient ?
  - Kudos to Vivek De (Intel) keynote on near threshold computing at the 2012 Lighter Than Green Dependable Multicore Architectures Workshop

Summary: Three Questions

- What are the right models to abstractly reason about energy, temperature and power as computational resources?



- What are the right models of processor heterogeneity and how does one manage/schedule heterogeneous processors?



- If computation at the lowest level is more error prone, how should higher levels be designed so that they are more resilient ?