The Run Time System and the End of Moore's Law: Why now? By who?

Yale Patt
The University of Texas at Austin

RoMoL
UPC/BSC
March 17, 2016

- ***Why now?***
  - ***We are used to continuing to get performance***
    - ***The DAC paper (55x): 7x from Moore; 55/7 from ILP***
  - ***Moore's Law may not be here much longer***
    (and by the way, when it goes, it will leave 50B transistors)
  - ***Can we continue to get more performance?***
- ***By who?***
  - ***By everybody!***

Optimizing resources at run time

- ***Some of the work we have done***
  - ***MorphCore (Khubaib, IEEE/ACM Micro 2012)***
  - ***Adaptive Processors, (Miftakhudinov, Micro 2012)***
  - ***Shared resources (IEEE/ACM ISCA 2012)***
  - ***Which thread to speed up (Joao, ISCA 2013)***
  - ***Enhanced memory controller (Hashemi, ISCA 2016)***
- ***All depend on knowing chip_resources/utilization***
- ***Can we do more?***

What we need in order to do more?

- ***Knowledge of the on-chip resources***
- ***Knowledge of their instantaneous utilization***
- ***Structures associated with the application***
- ***Close coupling of all of the above***

Chip Resources

- ***50 Billion transistors (some dark silicon)***
- ***Some writeable control store***
- ***FPGAs***
- ***Maxeler: a data flow graph***
- ***ASICs (FFT, GPU, Quantum connection)***
- ***Domain specific (we have reached the era of IoT!)***

Current Utlilization (Serious on-chip monitoring)

- ***Probably some SSMT microcode measuring activity***
- ***Perhaps some application specific SMT monitoring***

Structures associated with the application

- ***Bring back the PDP-11 EMT instruction***
- ***Architectural bit vector for on-chip FPGAs***
- ***On-chip SMT threads to help in monitoring***
- ***Simple pragmas associated with the code***
- ***Procedure call invocations for the run time system***
- ***OmpSs (What additional structures included)***

Close coupling

- ***The decision maker must be on the chip,***
- ***Aware of the on-chip resources,***
- ***Fed by the on-chip monitors, and***
- ***The structures passed down from above***

How do we accomplish all that?

- ***Programmers add the structures***
  - ***Perhaps a task graph, perhaps pragmas, whatever***
- ***Compilers insert procedure calls***
- ***Serious on-chip monitoring provides information***
- ***The Run-time system makes the timely decisions***

That is, …

Problem

Algorithm

Program

ISA (Instruction Set Arch)

Microarchitecture

Circuits

Electrons

Who understands this?

- ***Should this be part of students' parallelism education?***
- ***Where should it come in the curriculum?***
- ***Can students even understand these different layers?***

BUT that is a different lecture!

The point is, what we need is:

- ***Knowledge of the on-chip resources***
- ***Knowledge of their instantaneous utilization***
- ***A structure associated with the application***
- ***Close coupling of all of the above***

i.e.,

The Run time System at the End of Moore's Law

Thank you!