Scribe notes for Lecture 16 - Oct 23

McPAT: An Integrated Power, Area, and Timing Modeling Framework for Multicore and Manycore Architectures

1. This tool is mainly estimating

Power, Timing, Area

2. What is McPAT's general approach of estimating power/timing/area?

- Analytical model
- E.g. for timing estimation, analytical models of transistors and wires are used - resistance and capacitance obtained from the model and delay is calculated.
- This allows the researchers who don't fully understand the details of circuits and transistor to estimate the new architecture readily.
- In 1980s, computer architects aren't aware of circuits well.
- Then, CACTI 1.0 was the first one which enables researchers to explore the various cache designs with simple configuration change.

3. Initial Papers suggesting Analytical Models

Tomohisa Wada, Suresh Rajan, and Steven A. Przybylski, "An Analytical Access Time Model for On-Chip Cache Memories," IEEE Journal of Solid-State Circuits.

Johannes M. Mulder, Nhon T. Quach, and Michael J. Flynn, "An Area Model for On-Chip Memories and its Application," IEEE Journal of Solid-State Circuits.

- These works are collaborative works of architecture and circuit world. T. Wada et al. tried to model the cache memory with analytical models.
- However, since the model was targeting general cache architecture, the predictions might not be accurate for the custom cache architecture.

4. CACTI: Cache Access Time

- CACTI is written by Norm Jouppi who mainly has an architecture background.
- Their work is combining the pre-existing models and translate them to C language

- Architects can then estimate timings of various configurations of caches and compare the performance of each design.

- Back then, power is not important, so no power modeling exists in CACTI 1.0.

5. Accuracy of proposed Models

- Models innately has errors due to the abstraction.

- Then, the question is how accurate it is. Is it reasonably accurate for researchers to use?

- Equations (analytical model) itself is inaccurate. They should be validated how much errors they have.

- How to validate? Generating the results (AREA/Timing) using SPICE and compare with the model output.

6. Power estimation

- Gate/Transistor Level

- SPICE is precise but takes a lot of time

- Initial modeling work from Berkeley – Power Mill (Synopsys)

- EDA vendors now (Cadence, Magma) have built in power estimation tool

7. Wattch by D. Brook

- Knowing power numbers is good, but how does it related to architecture?

- First paper which incorporates power model with architectural simulator

- Run binary on simulator, and every cycle the activity information feed into the power model and aggregates the power numbers.

- Targeting single processor only / No static power since it was not critical in 2001.

- Blocks in the microarchitecture are analytically modeled : Reg File/CAM/ Result Bus, and etc.

- Wattch uses CACTI (memory models), and other works (adders, INT/FP FU).

- Starts from 18um process technology => lots of extrapolation.

- No timing or area estimation – only power

8. McPAT

- Integrating CACTI / Wattch / Orion

- Orion is interconnect model – wires are big problem now (this allows many/multi core system estimation).

9. Power

$P_{avg} = P_{switcing} + P_{shortcircuit} + P_{leakage} + P_{static}$

Dynamic power: $P_{switcing} + P_{shortcircuit}$

$P_{switching}$: αCV^2F (α is activity factor), $P_{shortcircuit}$: short circuit current when Fraction of time when both on.

- Static Power: Many types of leakage exist. Major leakage is sub-threshold leakage (I_{sub})
- Static power increase over the decades. However, new technology (SOI, high K metal gates, FinFET..) alleviates the problem.
- McPAT abstract away many types of the static current.

10. Metric in McPAT

- EDP: energy delay product, ED^2P: when performance is important
- EDAP: energy delay area product / EDA^2P: when cares about die cost more
- Tradeoff between power and delay: when more energy is used, delay decreases or the other way.

11. Multicore Performance Evaluation

- Power density decreases as the number of cores increases.
- EDP is best for 8 cores (if you don't worry about area)
- EDAP increases (from 4-cores to 8-cores) - questionable.

12. McPAT: Pros and Cons

- McPAT requires the activity information as inputs, but they innately has errors.
- McPAT models general architecture - specialized units are not modeled. (>= 60%)
- Hierarchical modeling is good.
- McPAT enabled computer architect to explore various systems without know the details of circuits.
- In general, McPAT may generate similar power results. However, there are chances that one error offsets the other in a good way.

- High level of core/system design. Fine when high level exploration, but in-depth research may require more accuracy.