Agenda

1 Introduction to myself

- Senior ASIC designer and lead engineer.
- Expert on Formal methods.
- In-depth and up-to-date knowledge of computer architecture and operating system.
- Leadership
  - Part of my education.
  - Part of my day-to-day work as lead engineer in chip projects.
- Vision into the future based on

2 Next generation ARM cores or processors for HPC and server

- Power wall → Dark silicon
- Process improvement → Cheap silicon
- Both → cores customized for special application

- ARM core + lots of accelerator cores for
  - Graph, DB,CFD,FFT, DNN…
  - Apple may want ARM+DNN for its Siri service
  - Facebook may want ARM+Graph for its social media service
  - Alibaba may want ARM+DB for its database intensive application
  - NUDT may want ARM+CFD+FFT for its HPC application
  - Amazon may want all for its cloud service

- A chip is partitioned into a grid of cells. Each cell contains
  - A TSV to 3D memory stacked on its top
  - An ARM core, or
  - A set of Graph, DB,CFD,FFT, DNN cores
  - These customized core may share some common infrastructure, such as
    - memory controller
    - floating point ALU

- So ARM research may become pretty busy in
  - Collecting or even forecasting customer application requirement
  - Proposing new architecture and software stack for these requirements
  - Profiling and fine tuning

- Benefits:
  - Expanding instead of changing ARM's current licensing business model.
  - Not affecting the research and development effort on ARM cores
  - May lead to a CoreStore similar to Apple's AppStore,
    - lots of third party accelerators designed by customers
    - ARM get lots of cores without direct investment.
  - ARM may capture share in application software market, such as database, HPC, by :
    - Developing software stack that can fully efficiently use its core
    - Providing consulting service to develop such software

Shortcoming: Too large a chip with many useless cores. But I don't think this is a problem:
Accelerator provides orders of magnitude improvement on performance and power, significantly offset the unused silicon area.

Server processors have very large margin in their price.

3 What to do in HPC and server market

Formally verifying complex architecture: modern high performance architecture include complex memory system with transaction memory, memory ordering and cache coherence. Formally verifying such complex architecture is a pretty hot topic recently:

MICRO15: CCICheck: Using µhb Graphs to Verify the Coherence-Consistency Interface
MICRO 10 : Fractal Coherence: Scalably Verifiable Cache Coherence
POPL 14 : Herding cats:Modelling,simulation, testing,and data-mining for weak memory
PLDI 15: Verifying read-copy-update in a logic for weak memory
ISCA15: ArMOR: Defending Against Memory Consistency Model Mismatches in Heterogeneous Architectures

Sketching method automatically filling details into template.

Normally, we design a complex structure with :

Some high level properties that can fully characterized its boundary behavior.
And some major structure decision such as pipeline stage, cache line replacement strategy.

In many cases, these relative simple specification can already determine almost all other details that can lead to correct design. And all other details can be filled in automatically by sketching.

PLDI 13 : TRANSIT: specifying protocols with concolic snippets
This one is tightly related to ARM's by synthesizing a cache coherence protocol.
CAV15 ： Adaptive Concretization for Parallel Program Synthesis
ASPLOS13 : Stochastic superoptimization

New architecture exploiting algebra and logic structure , for example:

Stochastic computing : using long random sequence to represent number, and simple logic gate as multiplier and adder. Very efficient in both power and area for media related workload, such as audio, picture and video processing.

DAC13 : Stochastic circuits for real-time image-processing applications
DAC15: Introduction to Stochastic Computing and its Challenges

Approximated computation: Trading off accuracy to achieve better performance and energy efficiency

PLDI 15: Automatically Improving Accuracy for Floating Point Expressions
MICRO 15: Doppelganger: A Cache for Approximate Computing
DAC 15: Joint precision optimization and high level synthesis for approximate computing
ISCA15: Rumba: An Online Quality Management System for Approximate Computing

Computing cores and hardware mechanism customized for special application

ISCA15 : A Scalable Processing-in-Memory Accelerator for Parallel Graph Processing
ASPLOS14: Integrated 3D-Stacked Server Designs for Increasing Physical Density of Key-Value Stores
ISCA13: LINQits: Big Data on Little Clients
ISCA15: Reducing World Switches in Virtualized Environment with Flexible Cross-world Calls
ISCA15: Accelerating Asynchronous Programs through Event Sneak Peek
MICRO15: Neural Acceleration for GPU Throughput Processors
MICRO 15: Fast Support for Unstructured Data Processing: the Unified Automata Processor

4 Opportunity in IOT market

Character:

- Low performance
- Unstable power
- Very large volume, maybe billions each year
- Very low end process, may be still 90nm.
- Very low price.
- Very simple chip structure and schematic

Observation:

Unstable power → Non-violated architecture and energy harvest

DAC15: Ambient Energy Harvesting Nonvolatile Processors: From Circuit to System

Unstable power → EDA algorithm for minimizing the state set to be saved

DAC15: Scalable sequence-constrained retention register minimization in power gating design

Low performance → Approximate computing mentioned above

4 Organization

Small number of large teams, focus on some relatively well defined problems.
Large number of small teams, free research, similar to an university.
Or something between.

Cooperation with universities and industrial partners

- Working with talent PIs and teams in Chinese universities.
- Working with large device manufacturer and users.
- Young intern from universities.

Diversity, may need expert from many very diverse field, not just architecture:

- Big data analysis
- Bioinformation
- Database
- Operating system
- Compiler
- Logic and formal method

Good scientist and engineer should be able to adapt to new problems by self-teaching.

6 Leadership and mentoring

Different leadership style

- Command and obedience.
    - Team member without active attitude
    - Leading to hot-spot in organization
- Free run within well defined boundary and exception handling
    - Team members with active attitude.
    - Already proved themselves in pursuing Master or PhD degree.
    - May be with some special habit, good for our diversity
    - Encouraging solving problem by self-teaching and self-organization
    - I only handle very urgent and difficult problems, or giving my opinion as senior.