Domain-Specific Architectures: The Next Wave of Computing Innovation

Antonio González

Director, ARCO Research Group
Professor, Computer Architecture Department

UPC, Barcelona, Spain

8th Workshop on Architectures and Systems for Big Data (ISCA 2018), Los Angeles CA, June 2, 2018

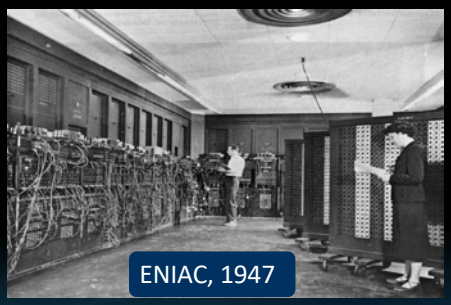
Agenda

- The next wave of innovation in computing
- Main hurdles and research directions
- A case study: automatic speech recognition
- Concluding remarks

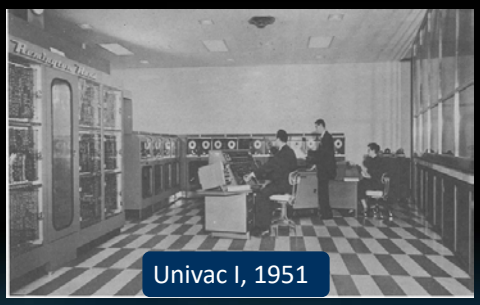
2

The First Revolution in Computing

The First Computers



ENIAC, 1947



Univac I, 1951

IBM 701, 1952

CDC 7600, 1969

3

The Second Revolution

Personal Computers

PC

Laptop

Ultrabook

Tablet

Convertible

Smartphone

4

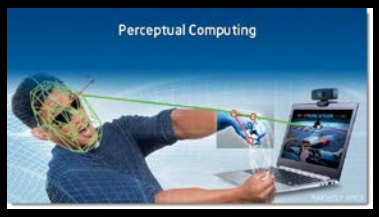
The Next Revolution: Ubiquitous Intelligent Computing

- Computing everywhere
  - On you
  - At home
  - At work
  - In the infrastructures
    - City
    - Roads
    - Public transportation
- Interconnected
  - To cooperate and share data
- Intelligent

5

Intelligent Computing

- Machines that can perform human-like intellectual tasks
- Some capabilities
  - Comprehend our surroundings
    - Vision
    - Language processing
  - Learn
  - Proactively take decisions and autonomous actions
    - Personal assistants
    - Drive assistants



6

Very Diverse in Functionality and Requirements

- Worn devices
- Body sensors / prosthetics
- Driving assistants
- Home robots
- Healthcare devices
- Energy management
- Smart consumer electronics




7

Require High Performance

- Complex tasks
  - Pattern recognition
    - Objects in real scenes
    - Spoken words
    - Facial identities and expressions
    - Anomalies (e.g. potential hazards when driving)
  - Natural language processing
  - Image and audio processing
  - Decision making
  - Etc.

8

And Must be Extremely Energy-Efficiency

- Small wireless devices have very limited battery capacity
- Performance ("intelligence") is limited by energy-efficiency
  - System power = EnergyPerTask * TaskPerSecond
  - Given a fixed power budget,
    - EPT has to decrease at the same pace as TPS (performance) increases

Reducing EPT is the key for delivering increased performance

9

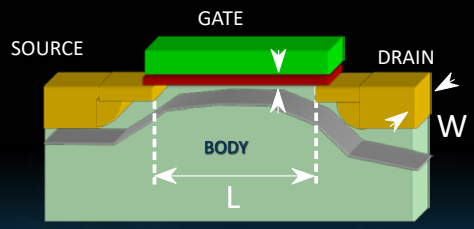
How To Improve EPT

- Technology
  - Scaling dimensions and other parameters

10

Ideal Technology Scaling

- Rules given by Robert Dennard in 1974
  - Scaling by a factor k
    - Device dimension (tox, L, W) 1/k
    - Doping concentration k
    - Voltage 1/k

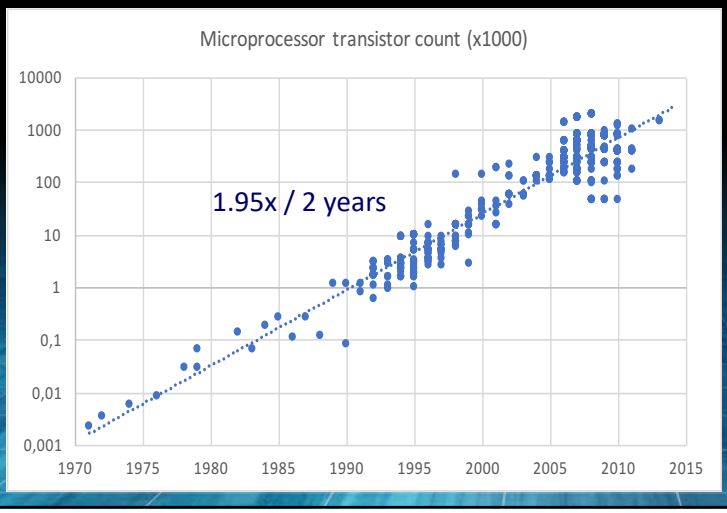


- Current 1/k
- Capacitance 1/k
- Delay time per circuit 1/k
- Power dissipation per circuit $1/k^2$
- Power density 1

11

Dimension Scaling Has Been Great

- Gordon Moore predicted doubling transistor density every 2 years (1975)



And it happened for the last 40 years

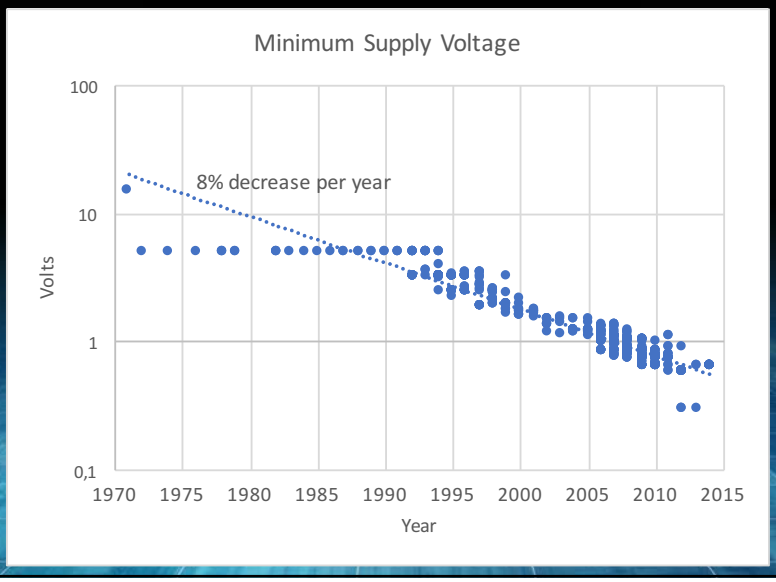
12

Future Projections

- Dimension scaling has recently slowed down
  - Previous 2-year cadence has increased to 3+ years
- May soon come to a halt
  - Silicon lattice spacing is 5.43 Å ≈ ½ nm
  - In two generations, dimensions will be around 10 atoms (10nm → 7 →5)

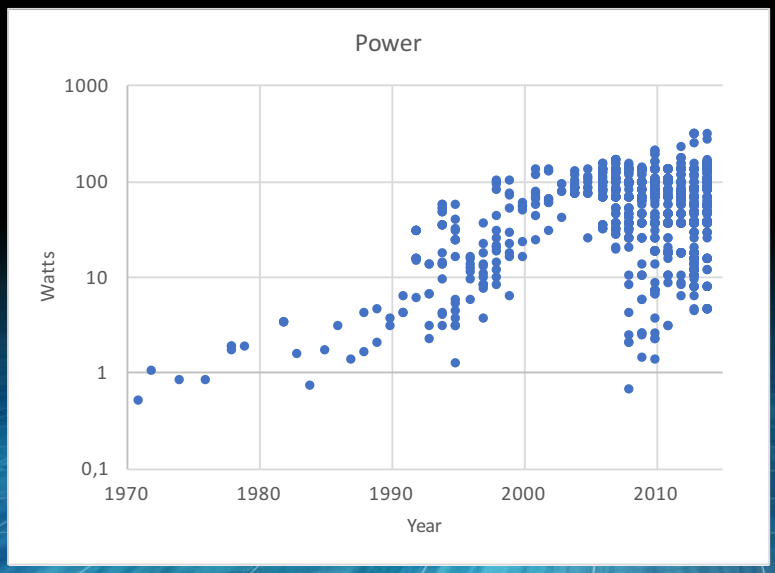
13

Supply Voltage Has Scaled Much Less Than Dennard's Rule



16

Consequence: Power Has Increased Over the Years

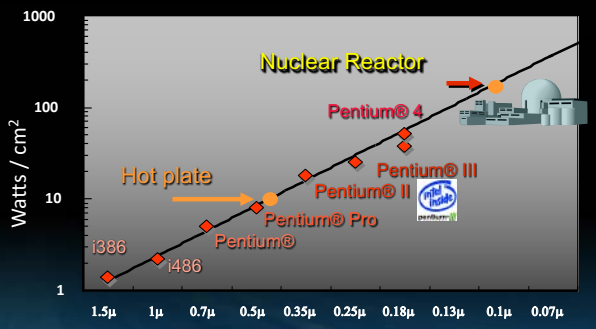


- Exponential increase
- Until early 2000s

17

Negative Effects of Power Increase

- Autonomy of mobile devices
- Cost of running the system
- Power density reached unsustainable levels
  - Cost of cooling solution
  - Form factor due to cooling solution
  - Noise of cooling solution
  - Reliability

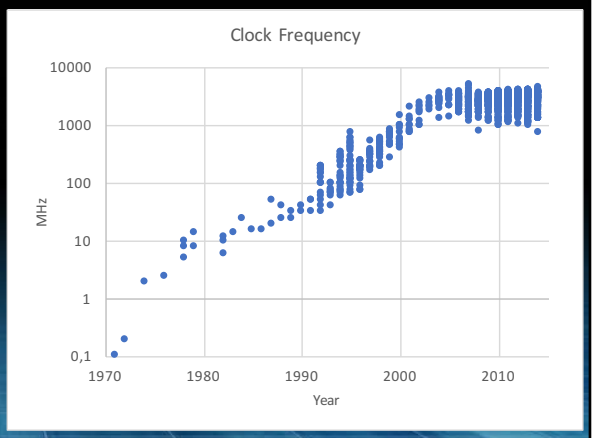


Source: S. Borkar, IEEE Micro 1999; F. Pollack, MICRO-1999; R. Ronen, WCED 2001

18

How Power Increase Was Stopped Around 2000

- Dynamic Power: C_{eff} x V_{dd}^2 x freq.
  - Clock gating
  - Reduces percentage of switching transistors
    - To reduce the effective capacitance
- Static Power: V_{dd} x I_{off}
  - Power gating
  - Reduces percentage of transistors that are powered on
- Sacrificing Performance
  - By not increasing frequency



19

How Power Increase Was Stopped Around 2000

- Dynamic Power: C_{eff} x V_{dd}^2 x freq.
  - Reducing the percentage of switching transistor
    - To reduce the effective capacitance
  - Extensive use of clock gating
- Static Power: V_{dd} x I_{off}
  - Reducing the percentage of transistors that are powered on
  - Power gating techniques
- Sacrificing performance
  - By not increasing frequency
- Specialized units
  - Specialization improves efficiency at the expense of flexibility

Qualcomm Snapdragon 820

Source: HotChips 2015

20

Future Technology Contribution To Improve EPT

- Summary
  - Scaling dimensions → Benefits may soon rich a point of diminishing returns
  - New technology → No mature alternative in the horizon

Innovations from architecture will be a key driving force in the forthcoming future

21

General Purpose Architectures

- Great improvements in the past
  - Pipelining
  - Caches
  - Branch prediction
  - RISC
  - Superscalar
  - Out of order execution
  - Multithreading
  - Multicore
- After 5 decades of evolution they are highly optimized and difficult to improve

Expected to provide minor improvements in the future

22

Time/Opportunity for Domain-Specific Architectures

- Key Features
  - Many simple units
    - Simple units have low performance but consume much less energy
    - Parallelism provides the desired performance at much lower energy cost
  - Much less data movement
    - For performance and energy reduction
  - More specialized hardware
    - Dramatic benefits in energy-efficiency
  - New ISA and programming paradigms
    - Oriented to "intelligence"-related tasks rather than numerical algebra

23

Example: Brain-Inspired Computing

- Human brain is very good at some of these intelligence-related tasks
  - E.g. object recognition
- Meets all key features described above
  - Composed of many simple units
  - Highly parallel
  - No centralized memory → only local data movements
  - With a very different programming paradigm: learning

24

A Case Study: Automatic Speech Recognition

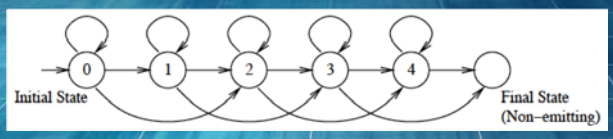
Large Vocabulary Continuous Speech Recognition

- A hard task
  - Word boundaries are not known in advance
  - Co-articulatory effects are very strong
  - Real time requirements
  - Low power constraints when carried out in mobile devices
- Main approach
  - Hybrid scheme
    - Hidden Markov Model + Artificial Neural Network

26

Modeling Speech

- Acoustic model
  - An utterance consists of a sequence of units of speech
  - Context dependent phones are the units used by most systems
    - About 50 phones in English
    - A phone does not always sound the same, due to co-articulatory effects
    - A triphone is a phone observed in the context of a preceding and succeeding phones
  - Each triphone is modeled by a small HMM (5-state in Sphinx II)



27

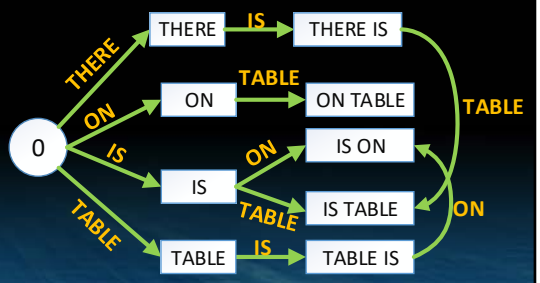
Modeling Speech

- Acoustic model (cont'd)
  - In the order of 1M or more states
    - Training data may be insufficient or too costly
    - States are often clustered (tied) into equivalence classes called senones
  - Word HMMs are built by concatenating the HMMs of individual triphones
  - There may be more than one model for a given word, to account for alternative pronunciations
  - Continuous speech is modeled by adding null transitions from the final state of every word to the initial state of all words in the vocabulary
    - Cross-word transition probabilities are given by the language model (e.g. trigram)

28

Modeling Speech

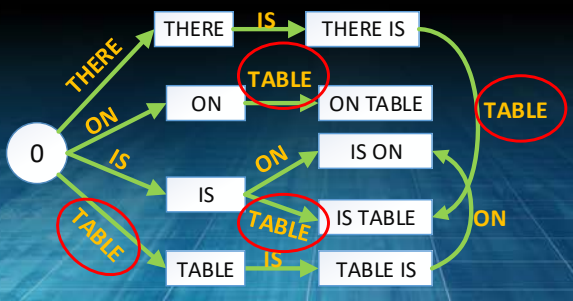
- Language model (aka grammar)
  - Helps to select the most likely word sequence from alternative hypotheses produced during search process
  - Example
    - Trigram: triples and their probability of occurrence
  - Backoff mechanism (example for trigram grammar)
    - Only the most frequent trigrams are included
    - If the desired trigram is not found, one falls back to bigram or unigram probabilities



29

Offline Composition

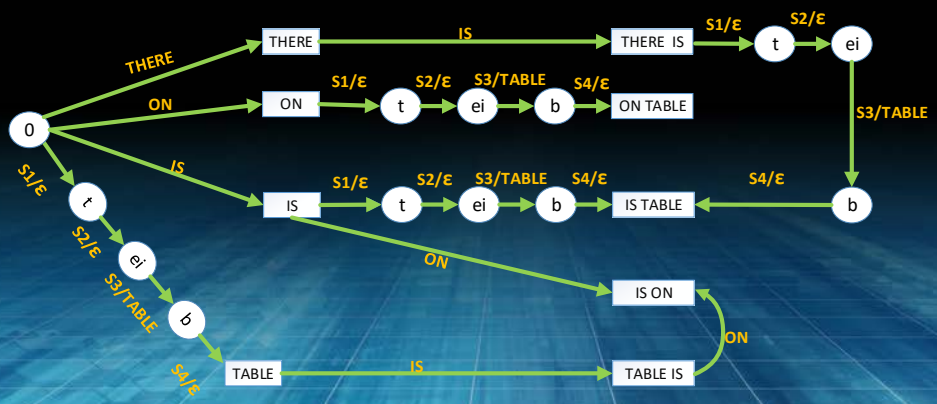
- In the language model, every word transition is replaced by its corresponding acoustic model



30

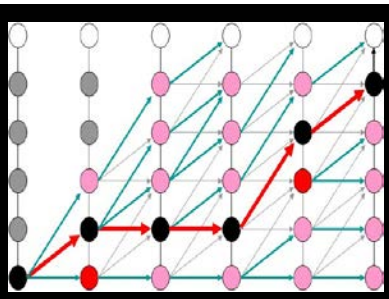
Offline Composition

- Resulting in huge WFST (e.g. 22M states, 1.1 GB for Kaldi Tedlium)



31

Viterbi Beam Search



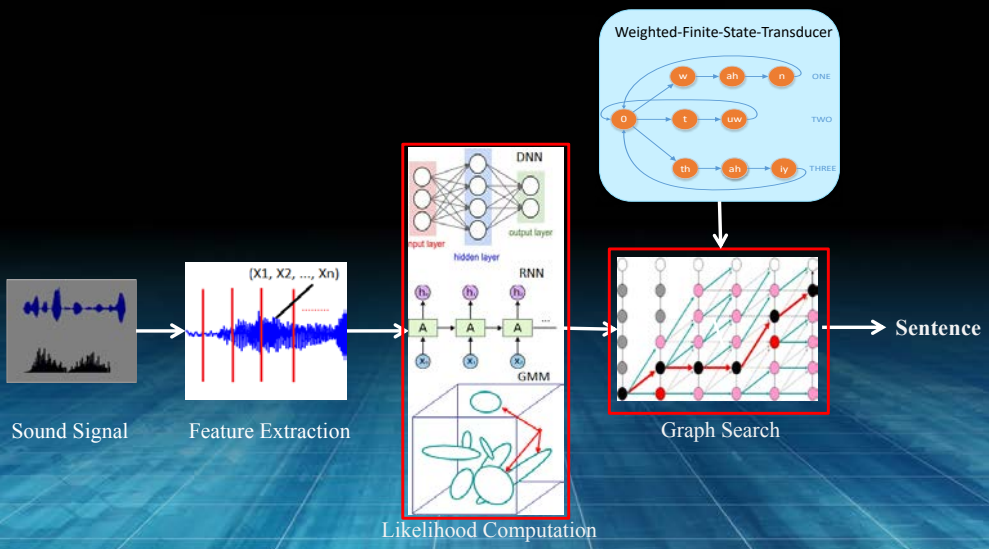
- Evaluates all transitions between previous frame and current one
- For every state it computes the probability corresponding to the best state sequence from time 0 to time t

$$P_j(t) = \max_i (P_i(t-1) \cdot a_{ij} \cdot b_i(t)), i \epsilon \text{ set of predecessor states of } j$$

- Each state has a single best predecessor, which is recorded
- The best state sequence is determined by starting at the final state and going backwards following the best predecessor at each point
- Beam search heuristic to reduce the search space
  - At each time t, the state with max probability is found: Pmax(t)
  - All states with probabilities lower than Pmax(t) x B are not considered for computations at time t+1
  - B is an appropriately chosen threshold lower than 1 that determines the width of the "beam"

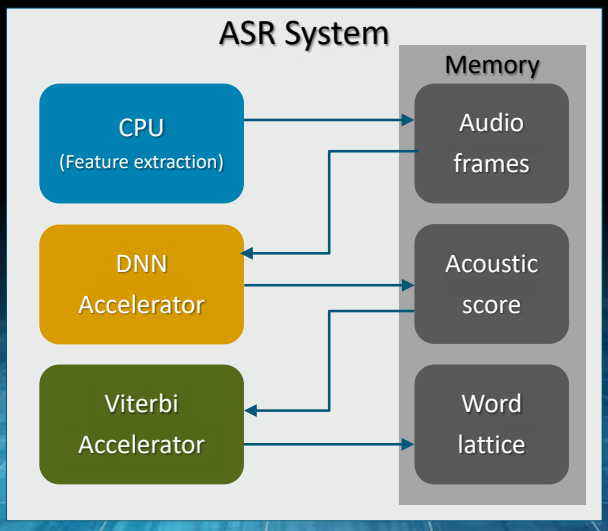
33

Putting It All Together



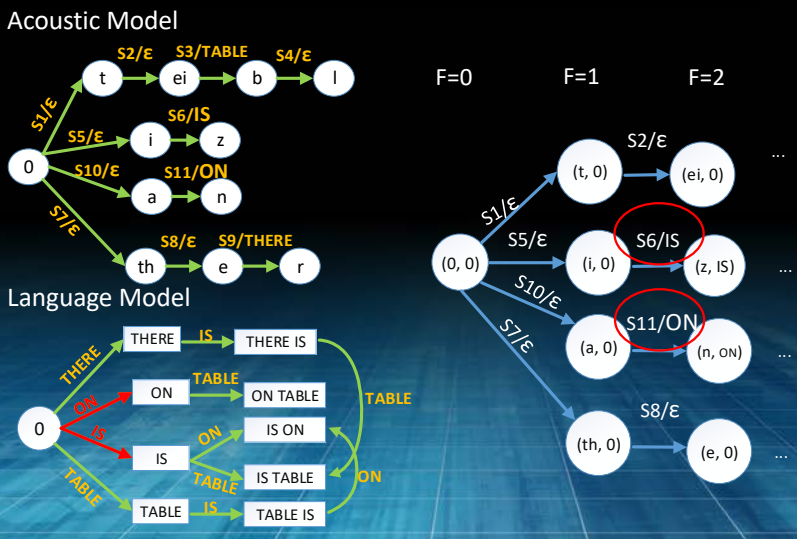
34

Accelerated ASR System

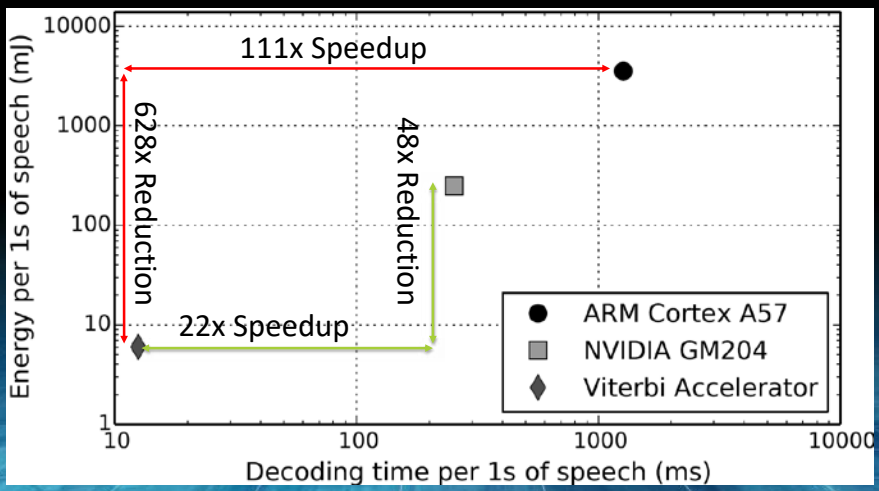


35

UNFOLD: On-The-Fly Composition [MICRO 2017]

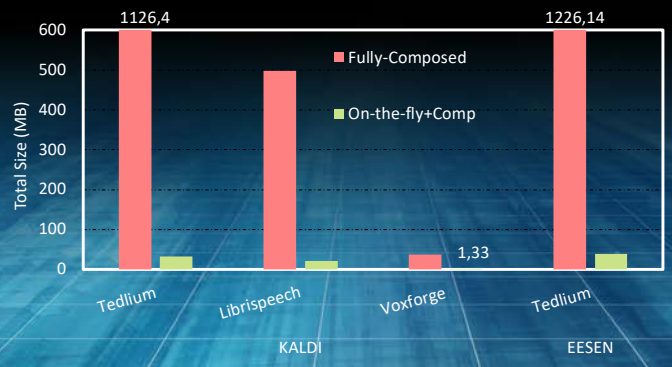


Viterbi Accelerator



Memory Requirements

- Huge reduction compared with fully-composed approach
  - 31x reduction in storage
  - 70% reduction in memory bandwidth



38

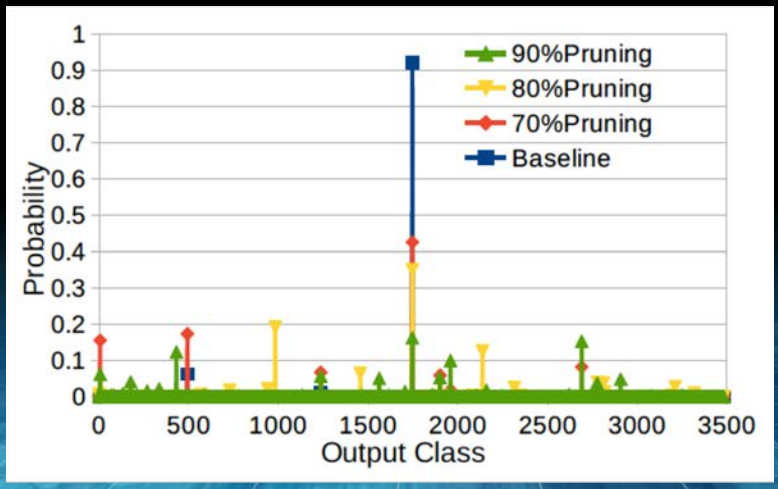
DNN Pruning [ISCA 2018]

- DNN pruning has recently become popular
  - DNN are usually oversized so many neurons and connections can be removed without impacting accuracy
  - Several recent works proposed different heuristics with high effectiveness (>50% pruning)
  - Large reduction in computations, energy and storage requirements
- However: How accuracy is measured?
  - Pruning studies normally use Top-1 or Top-5
  - This may not be adequate in some use cases of DNN such as ASR

39

Our Observation: DNN Pruning Reduces Confidence

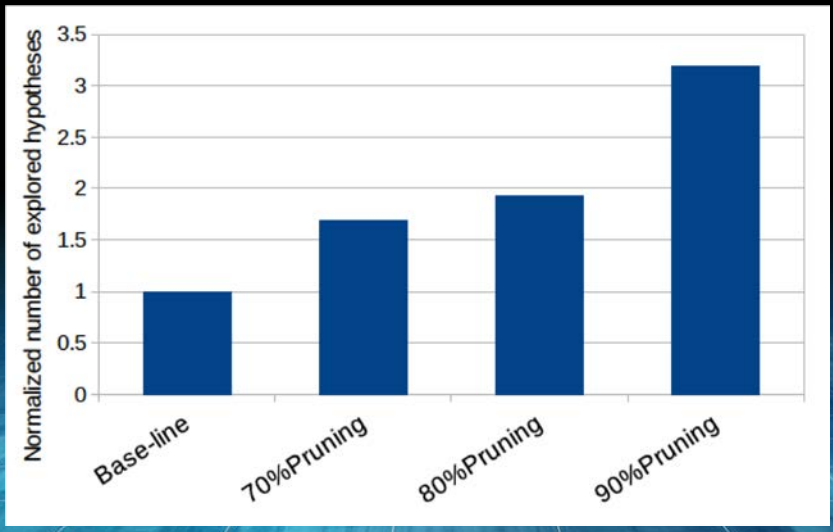
Acoustic scores for one particular frame



Top-1 is the same for all but its likelihood is reduced drastically

40

The Consequences: Less Confidence Requires to Evaluate More Hypothesis



41

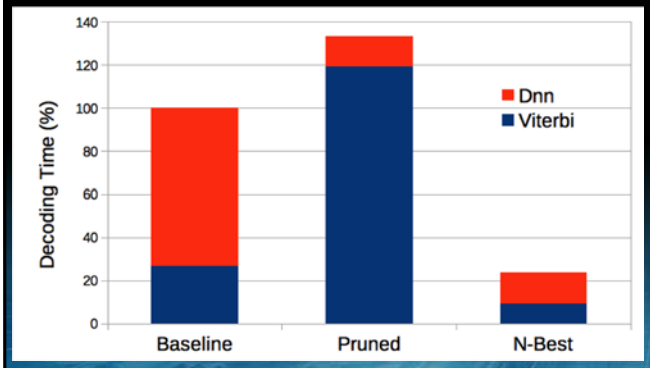
The Idea [ISCA 2018]

- Keep the N-best hypotheses, regardless of confidence
- The problem: Compute the N-best is costly
  - Requires sorting of a huge number of hypotheses (20K on average, up to 300K in Kaldi-LibriSpeech)
- Our proposal: Approximate the N-best
  - Using a K-way set-associative hash table
  - Keeping the K-best for each entry

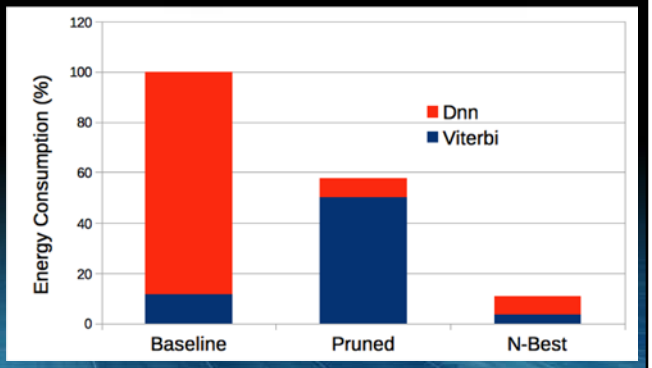
42

Performance and Energy Consumption

- 5.6x speedup



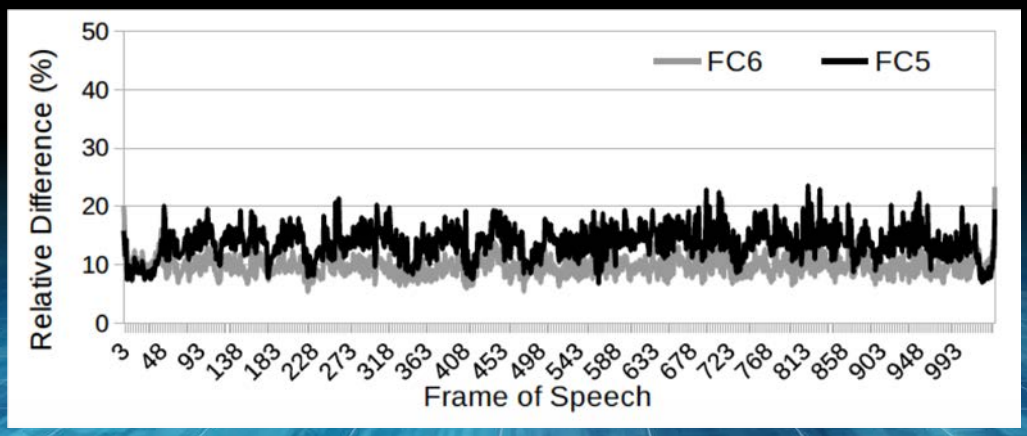
- 5.2x energy savings



43

Improving the DNN Through Computation Reuse

- Observation: Consecutive inputs are quite similar



44

Our Proposal: Computation Reuse [ISCA 2018]

- Reuse previous output
- Make adjustments to take into account input variations
  - If number of different inputs is small, adjustments are cheaper than computing the output from scratch

45

Results

- Input similarity: 45%
- Computations saved: 53%
- Speedup: 1.9x
- Energy reduction: 49%

46

Summary

- Next revolution in computing
  - A broad variety of intelligent devices
  - Ubiquitous
  - Applications very different to typical number crunching
  - Dramatic improvements in energy efficiency are required
- Call for domain-specific architectures
  - Massive parallelism
  - Reduction in data movement
  - More specialized hardware
  - New programming paradigms
- Case Study: UNFOLD, an efficient speech-recognizer for mobile system
  - 550x real-time (1.8 ms per second of speech)
  - 1.4 mW average power (1.4 mJ per second of speech)
  - 18 mm2 (11 for Viterbi + 7 for DNN)
  - Less than 30 MB of DRAM memory

47