A Scalable Speech Recognizer with Deep-Neural-Network Acoustic Models and Voice-Activated Power Gating

Michael Price*, James Glass, Anantha Chandrakasan

MIT, Cambridge, MA
* now at Analog Devices, Cambridge, MA

Diversification of speech interfaces





Today

- Personal assistant (smartphone)
- Search (smartphone, PC)

Future

- Wearables
- Appliances
- Robots

Goal:

Complement or replace touchscreen interfaces

Computational demands of ASR

(ASR = Automatic speech recognition)

- Real-time ASR is now feasible on x86 PCs
- ARM SoCs can run with minor performance degradation
- But what about everyone else?
  - Today's model: offload to cloud servers
  - Tomorrow: offload to cloud OR hardware accelerator

System power concerns:

- Memory
- I/O signaling

If processor efficiency is optimized in isolation, these can exceed core power

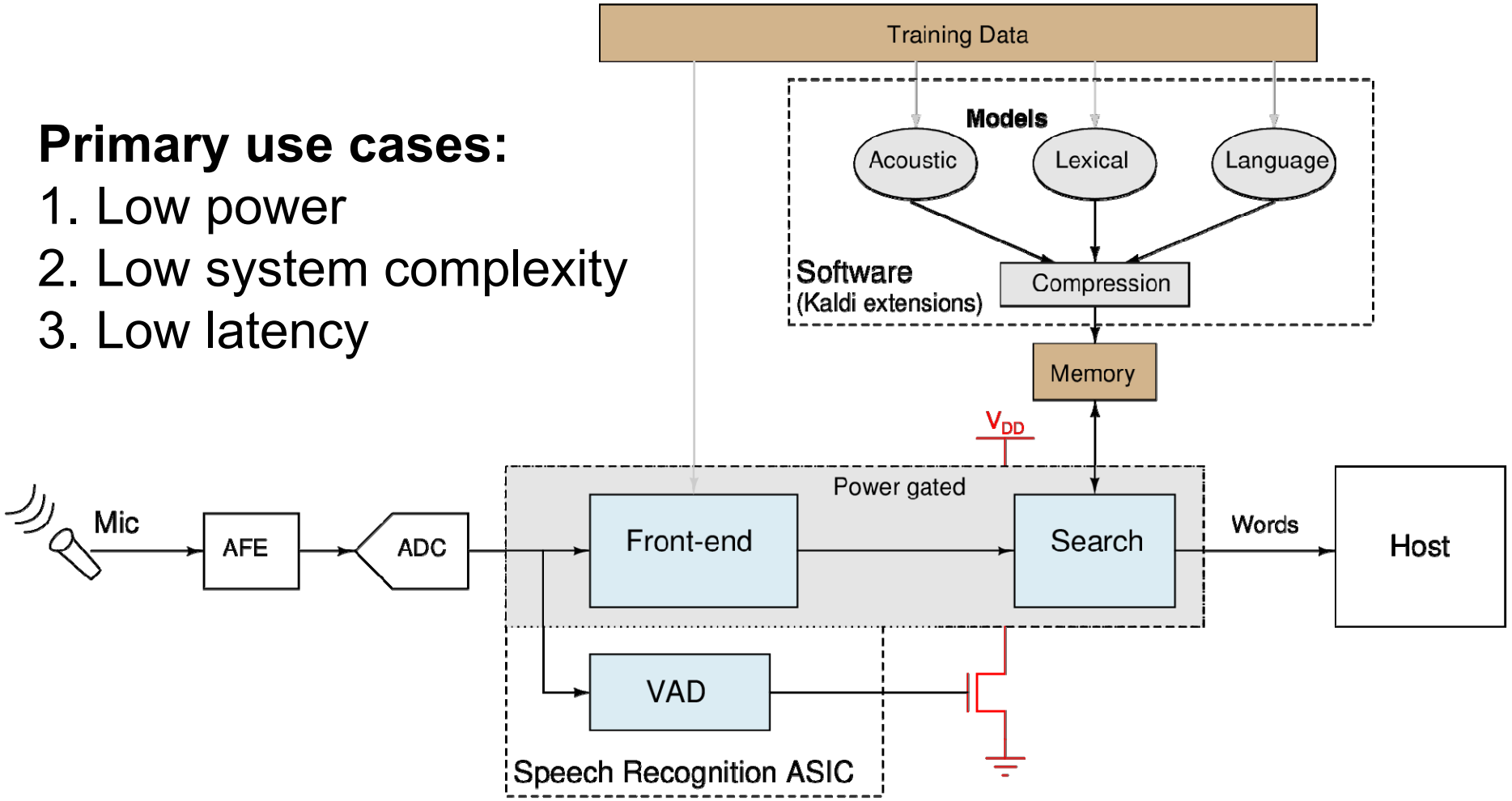
Today's non-volatile memory:
MLC NAND flash

At 100 pJ/bit: 100 MB/s -> 80 mW

Hardware accelerated speech interface

Primary use cases:

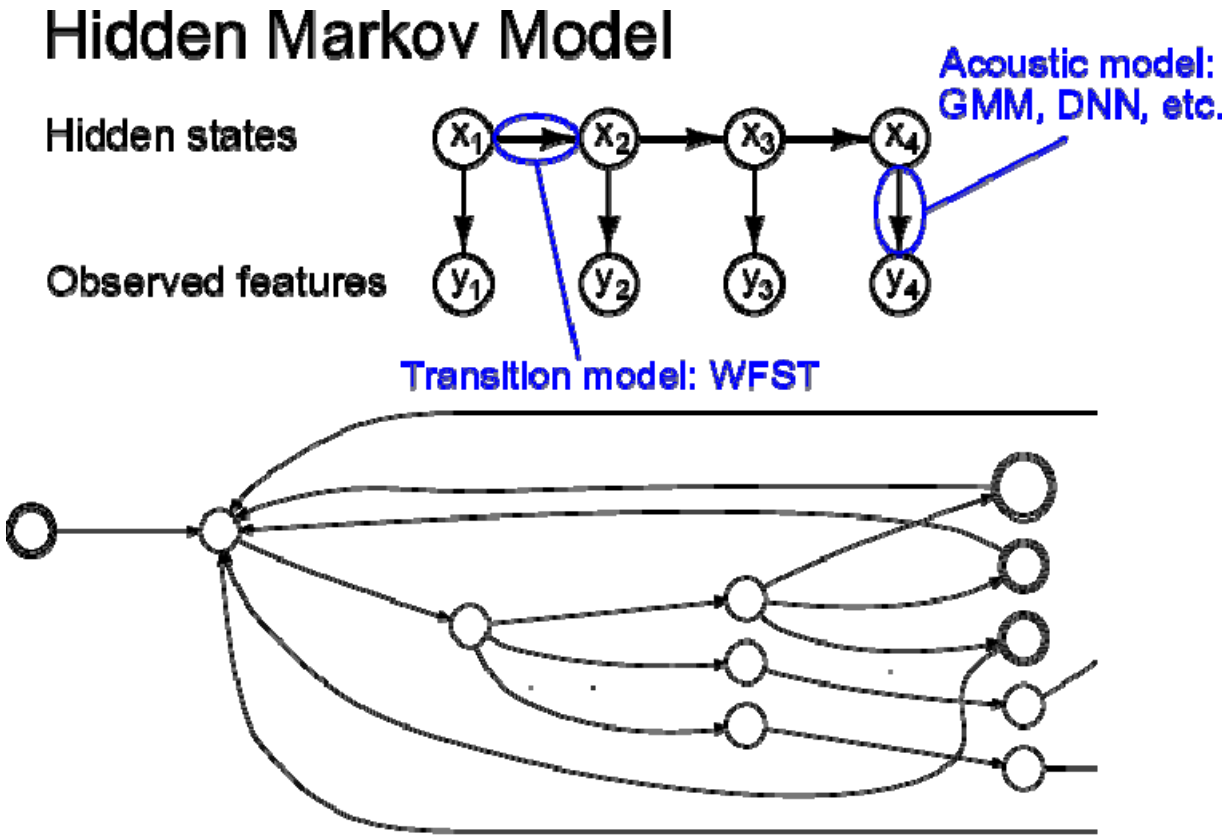
1. Low power
2. Low system complexity
3. Low latency



Outline

1. **Introduction**
   a) Motivation and scope
   b) ASR formulation
2. **Acoustic modeling with deep neural networks (DNN)**
   a) Performance with limited memory bandwidth
   b) Parallel architecture
   c) Details of execution unit design
3. **Search (Viterbi) architecture**
4. **Voice activity detection (VAD)**
   a) System power model
   b) Modulation frequency (MF) algorithm and architecture
5. **Test chip**
   a) Circuit features
   b) Measured performance

ASR formulation – HMM

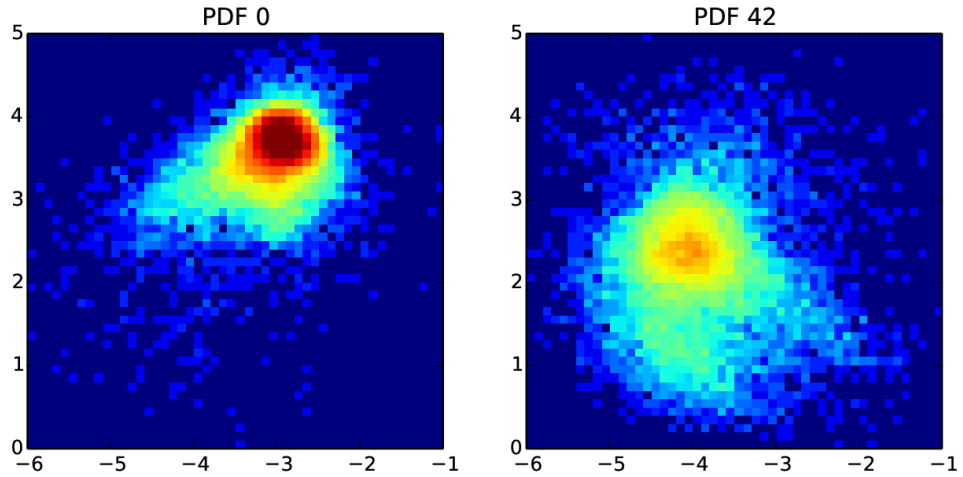


Inference: search using Viterbi algorithm

$$p(x_{t+1}) \approx \max_{x_t} p(x_t)\, p(x_{t+1}|x_t)\, p(y_{t+1}|x_{t+1})$$

ASR formulation – Acoustic model

- Training clusters WFST states into “tied states” (senones)
- Acoustic model approximates $p(\mathbf{y} \mid i)$
  where i is the tied state index and $\mathbf{y}$ is the feature vector (typ. 10—50 dims.)

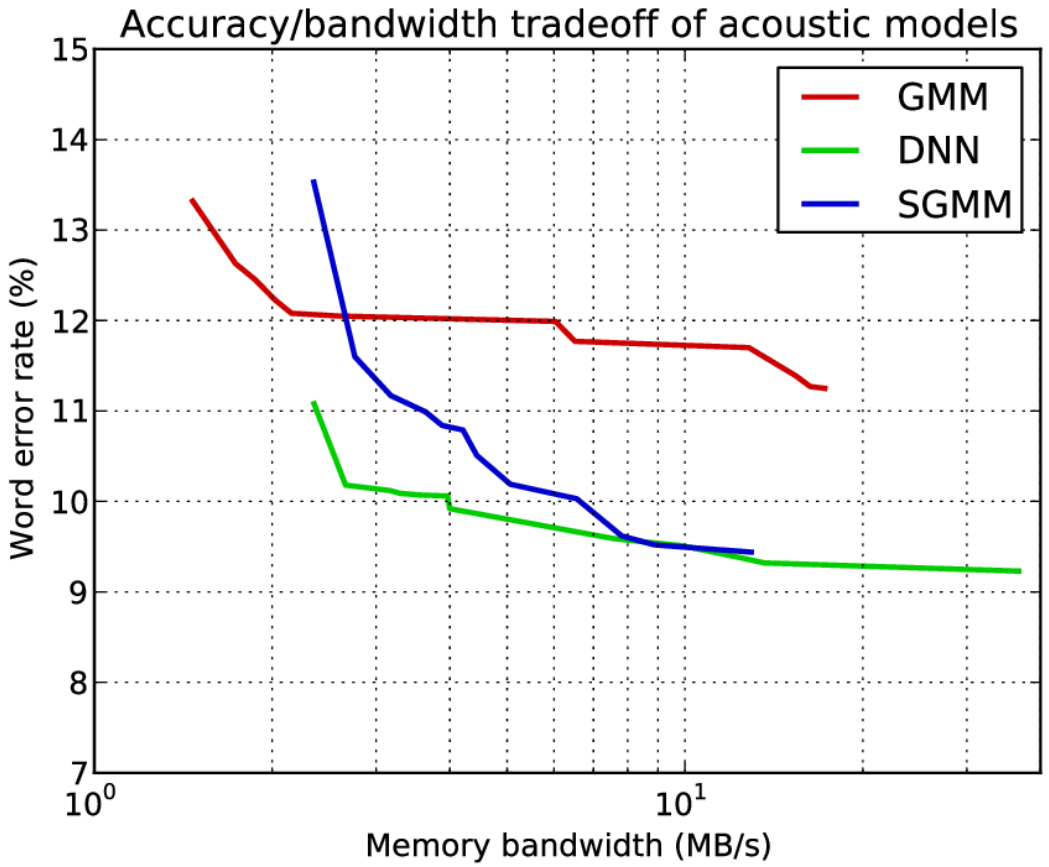


Example PDFs
(MFCC features
projected to 2-D)

Models have many parameters – large memory requirement

- **Size**: Typical ASR model is ~50 MB – must be stored off-chip
- **Bandwidth**: Naïve evaluation would require 5 GB/s
  (Target for low-power ASR: ~10 MB/s)

Bandwidth-limited acoustic models



Comparison of frameworks considers:

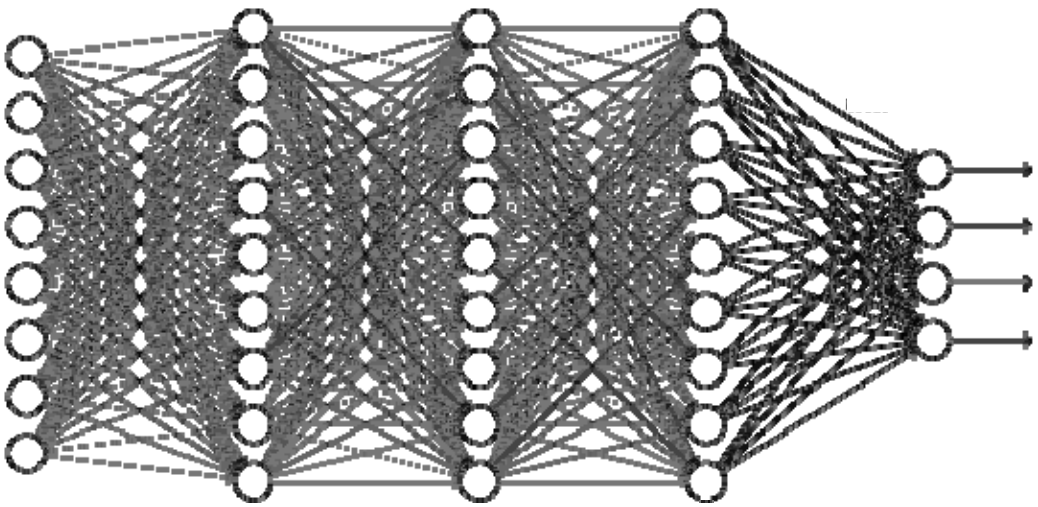
- Quantization
- Parallelization
- Accuracy

GMM = Gaussian mixture model

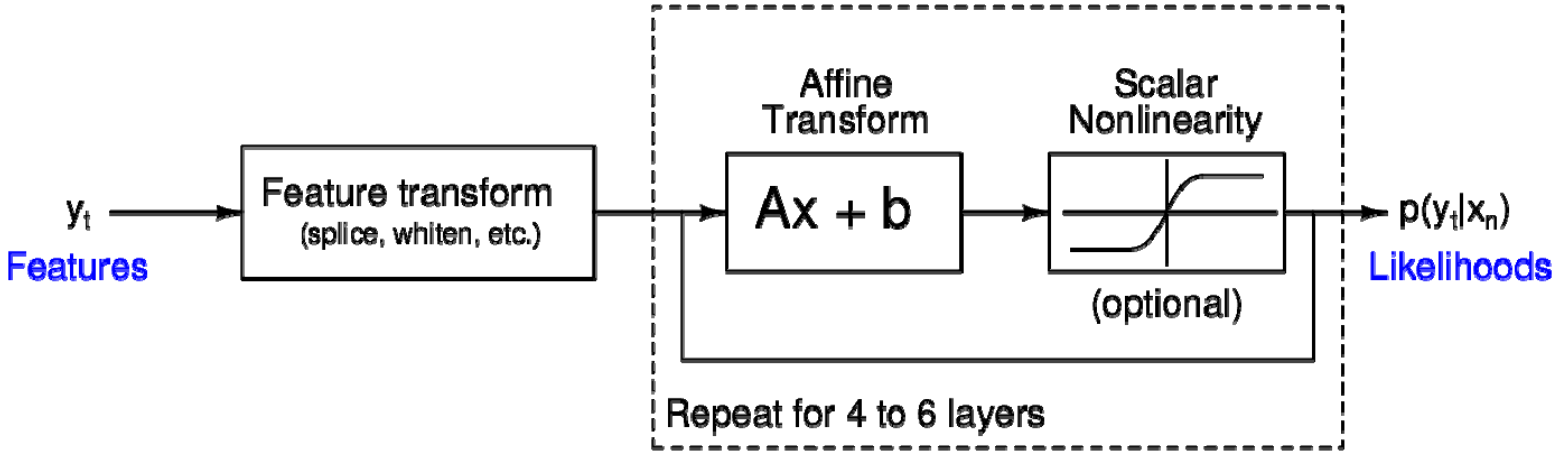
SGMM = Subspace GMM

DNN = Deep neural network

Feed-forward neural network (DNN)

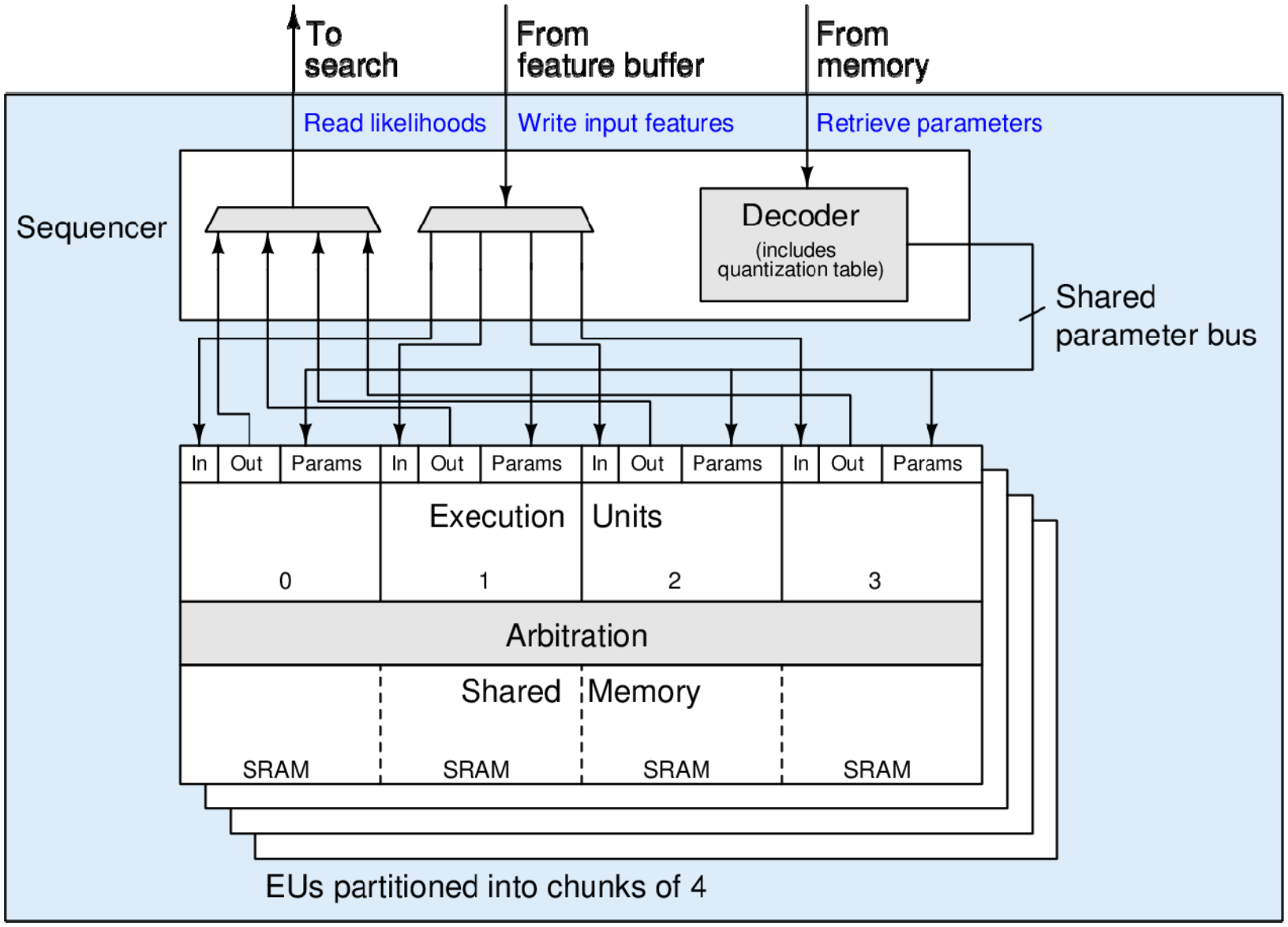


- Sequence of simple nonlinear transformations
- Used in ASR to evaluate all acoustic likelihoods jointly
- This work: **feed-forward networks only** (no time dependence)



Top image is from: Michael A. Nielsen, “Neural Networks and Deep Learning”, Determination Press, 2015.

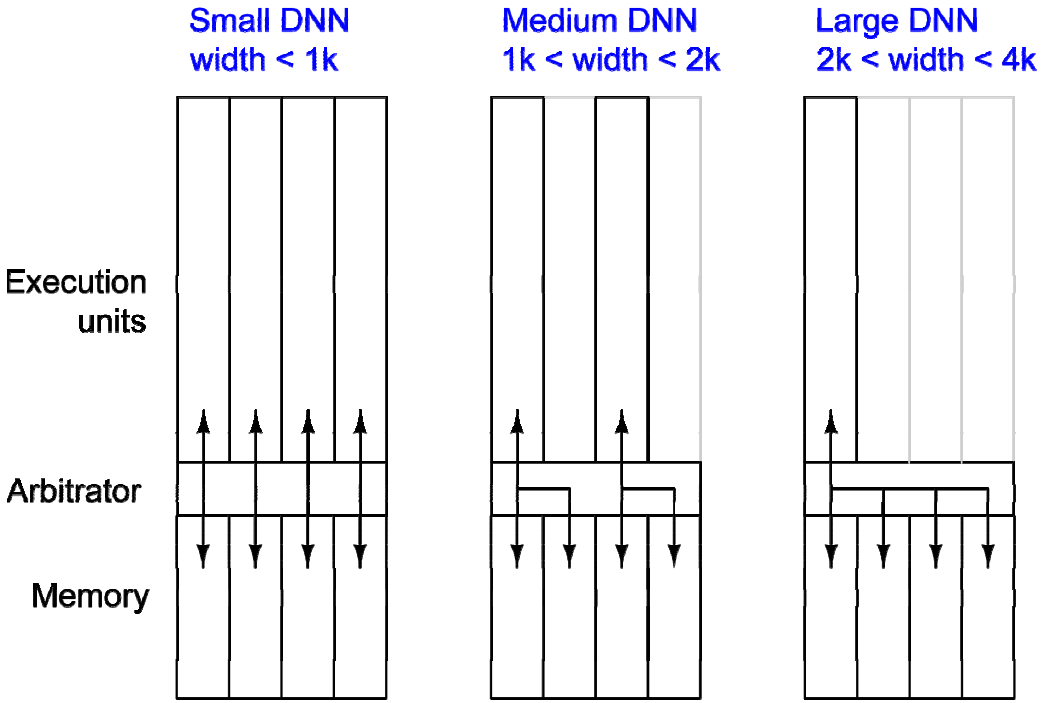
DNN evaluator architecture

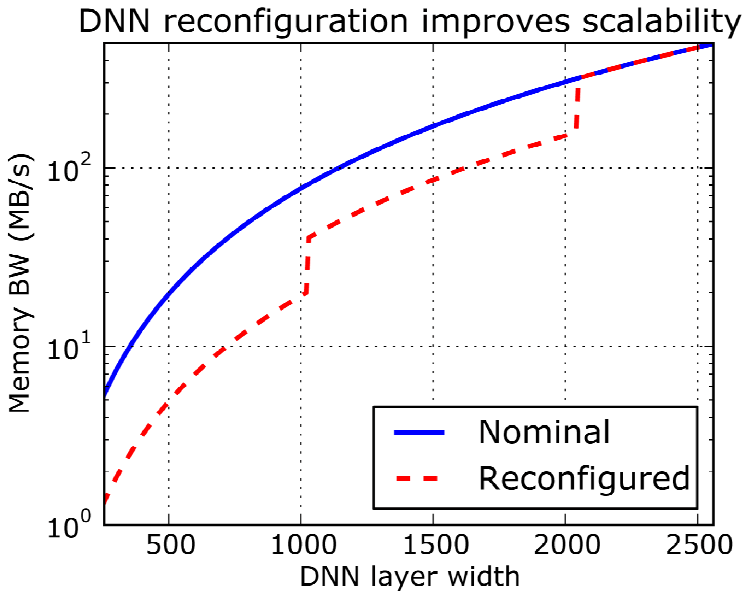


Execution unit assignment

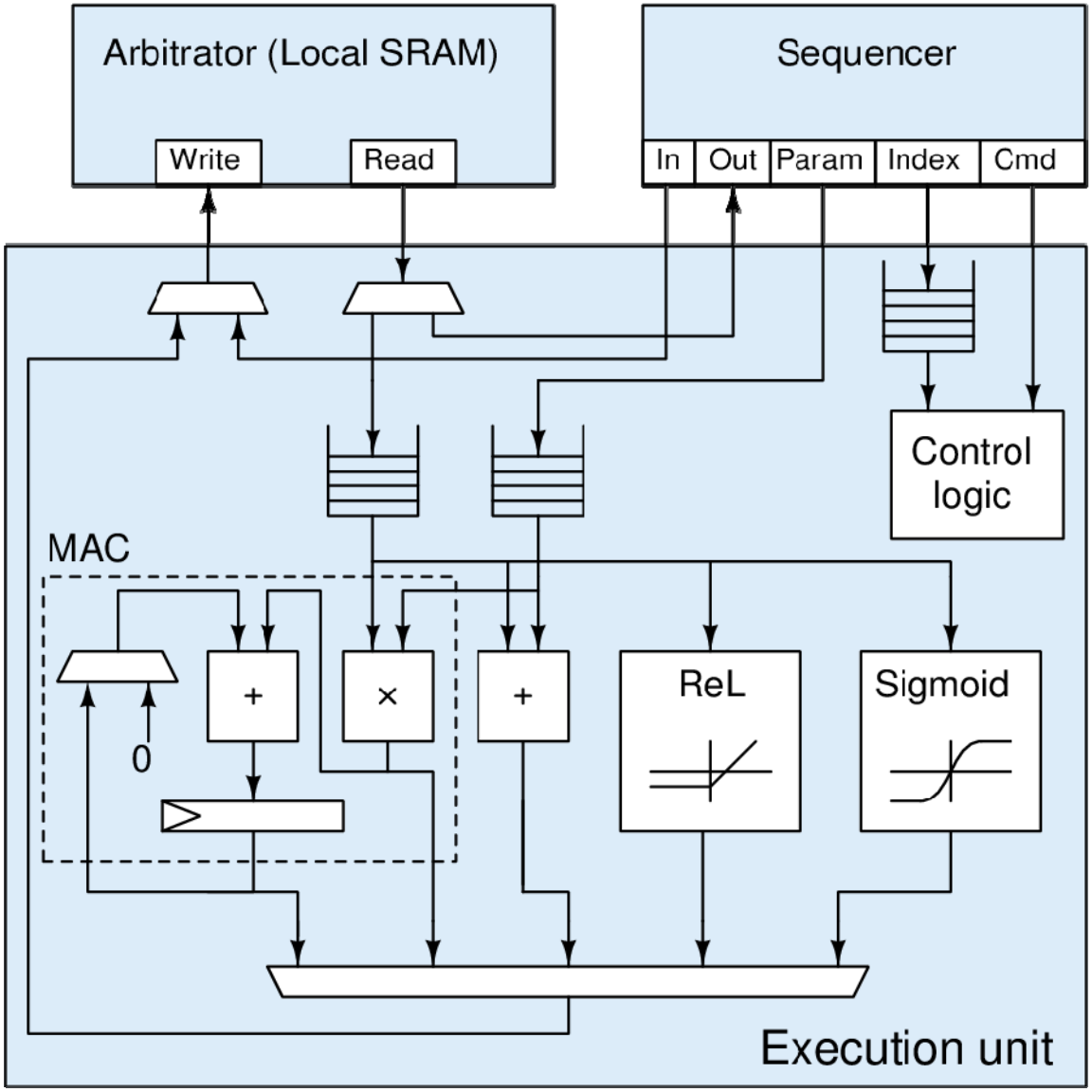
Configuration options for each DNN chunk (4 EUs)

Memory bandwidth scaling





Execution unit design



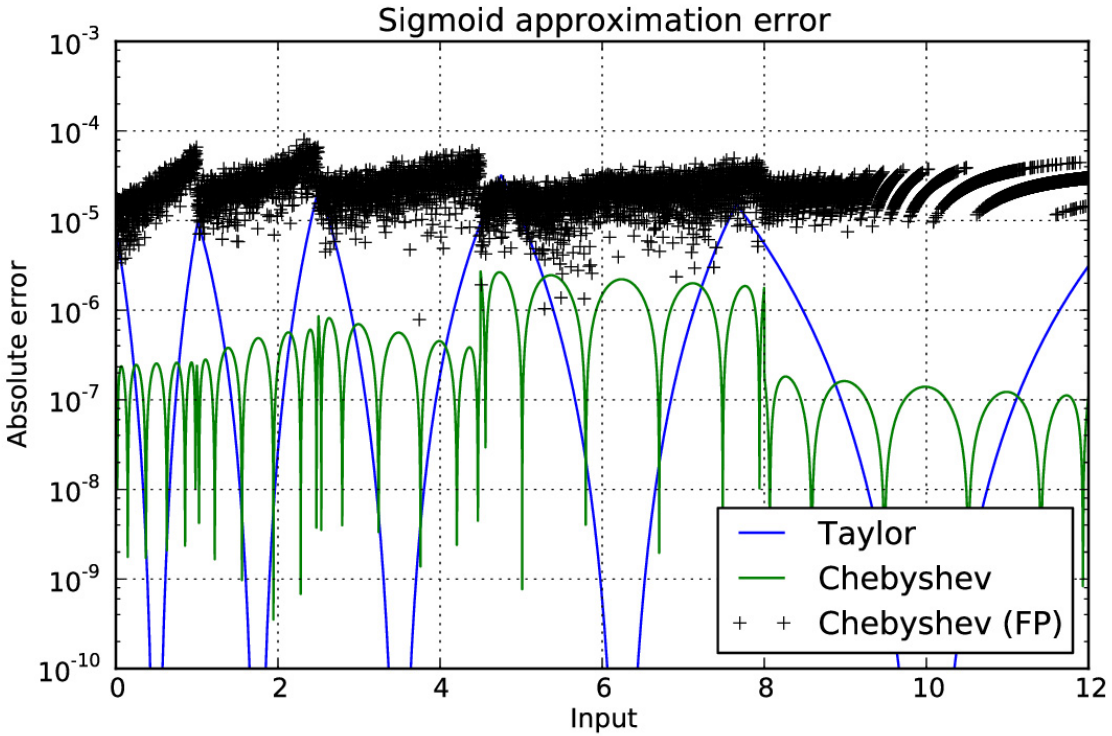
- **Complexity has been minimized:**
  - Executes arithmetic commands from sequencer
  - Writes results back to local SRAM

Sigmoid approximation

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Piecewise (low-order) polynomial fit

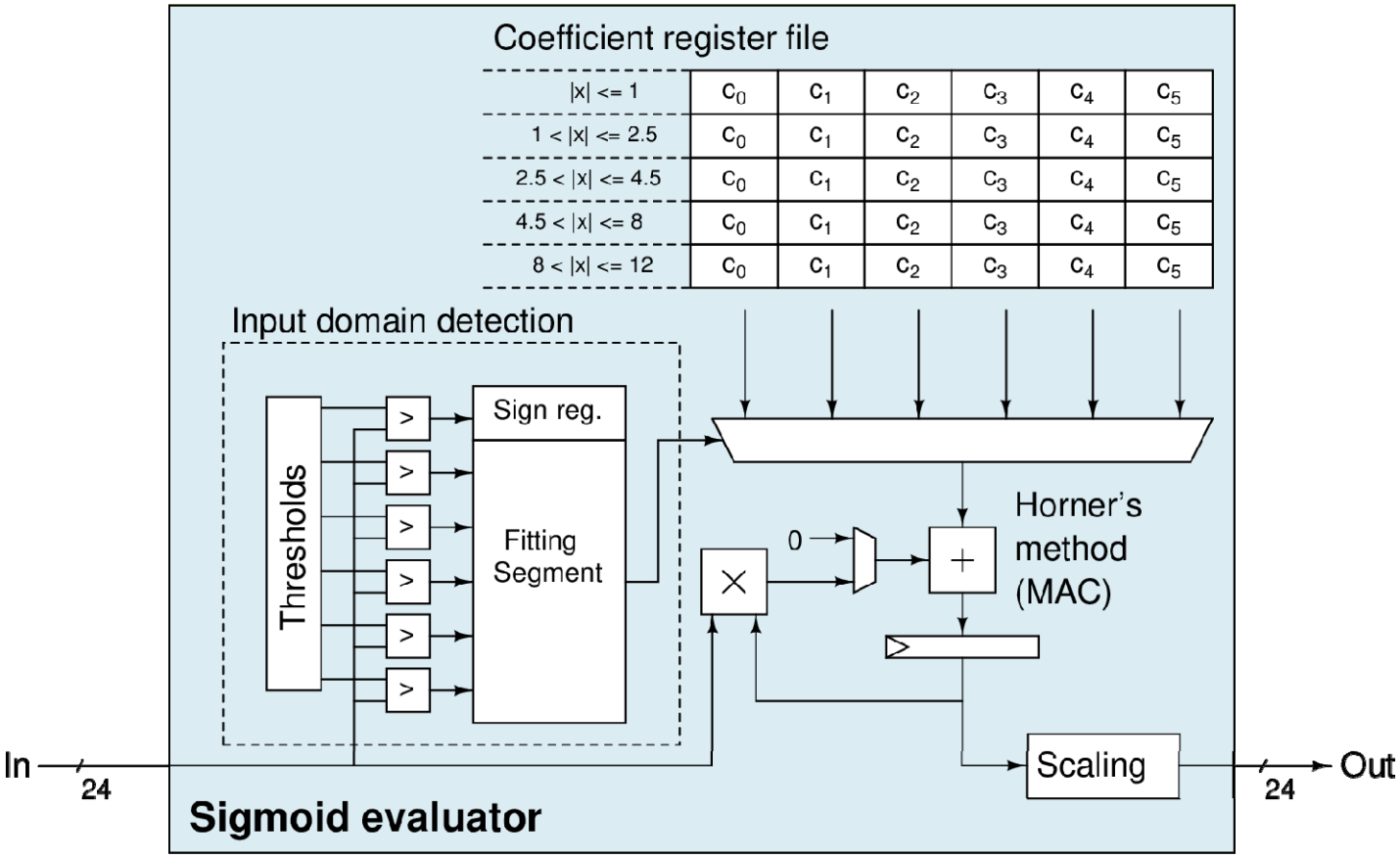
- Less area than plain LUT
- Simpler to evaluate than high-order fit



Chebyshev approximation

- Better accuracy than Taylor series

Sigmoid approximation

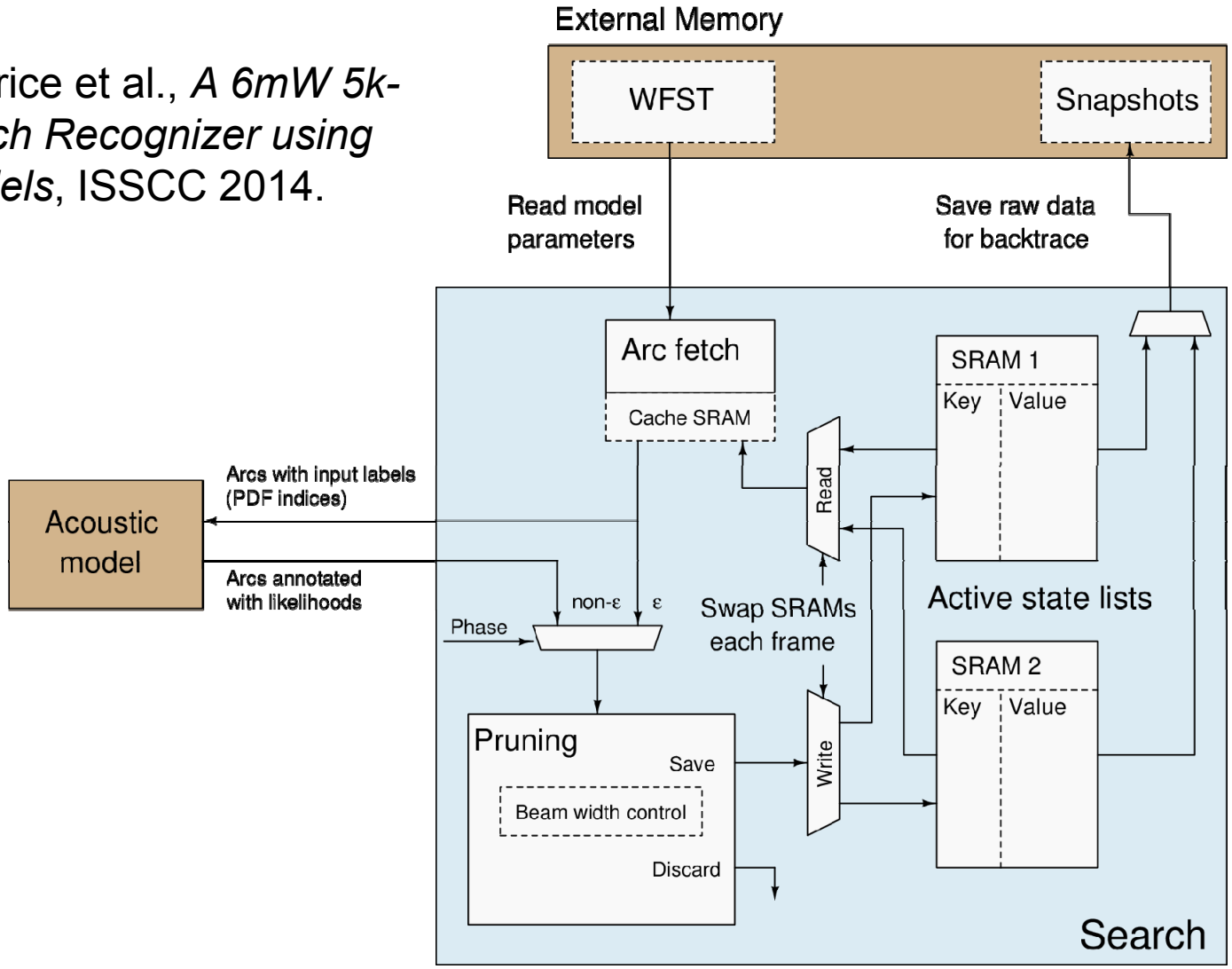


- Polynomial evaluated using Horner's method
- Use unpipelined version to save area
- 7-cycle latency; 5.7k gates (mostly the multiplier)

Search – Viterbi algorithm

Baseline architecture

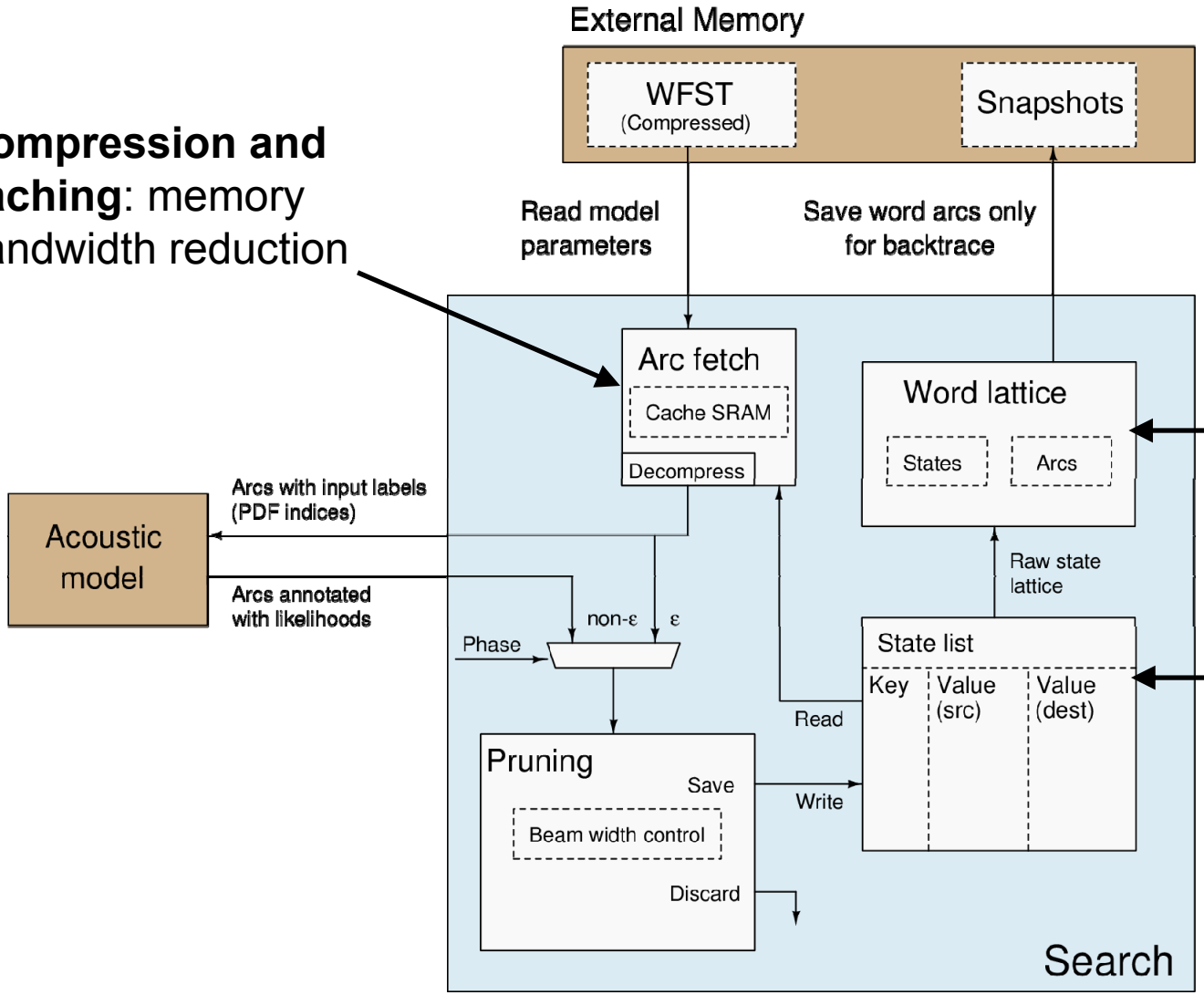
From: M. Price et al., *A 6mW 5k-word Speech Recognizer using WFST Models*, ISSCC 2014.



Search – Viterbi algorithm

Improved architecture

Compression and caching: memory bandwidth reduction

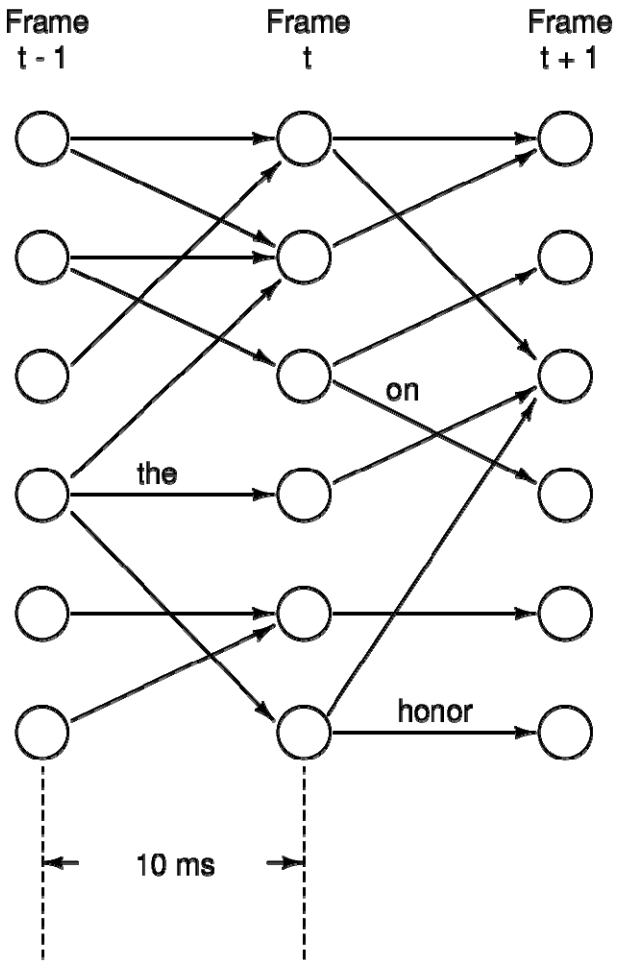


Word lattice: memory bandwidth reduction

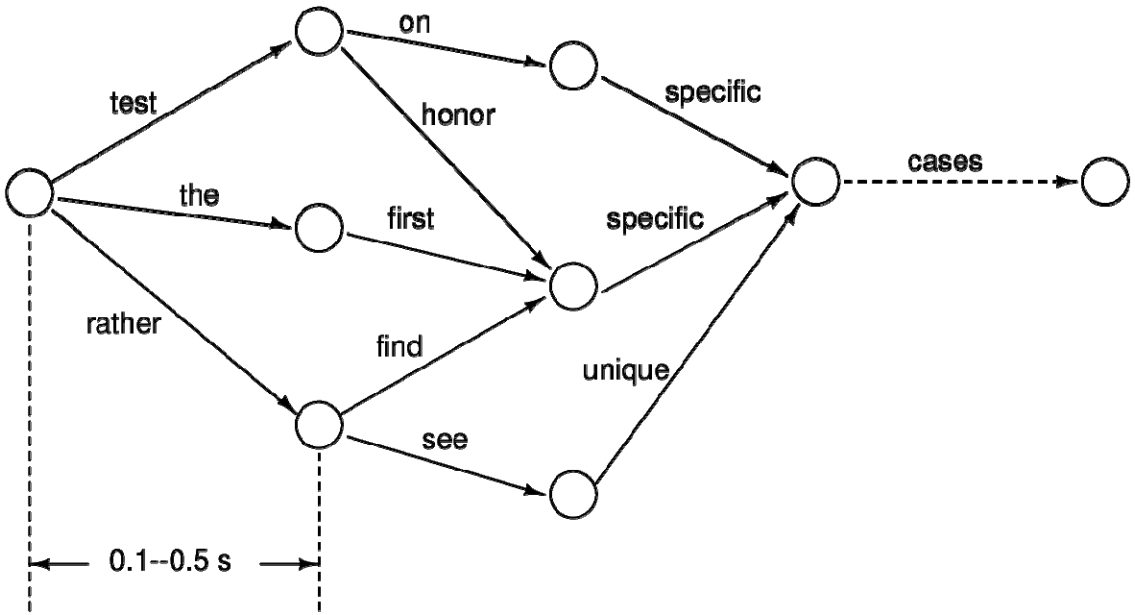
Merged state lists: 11% area savings

Search – Word lattice

From state lattice…

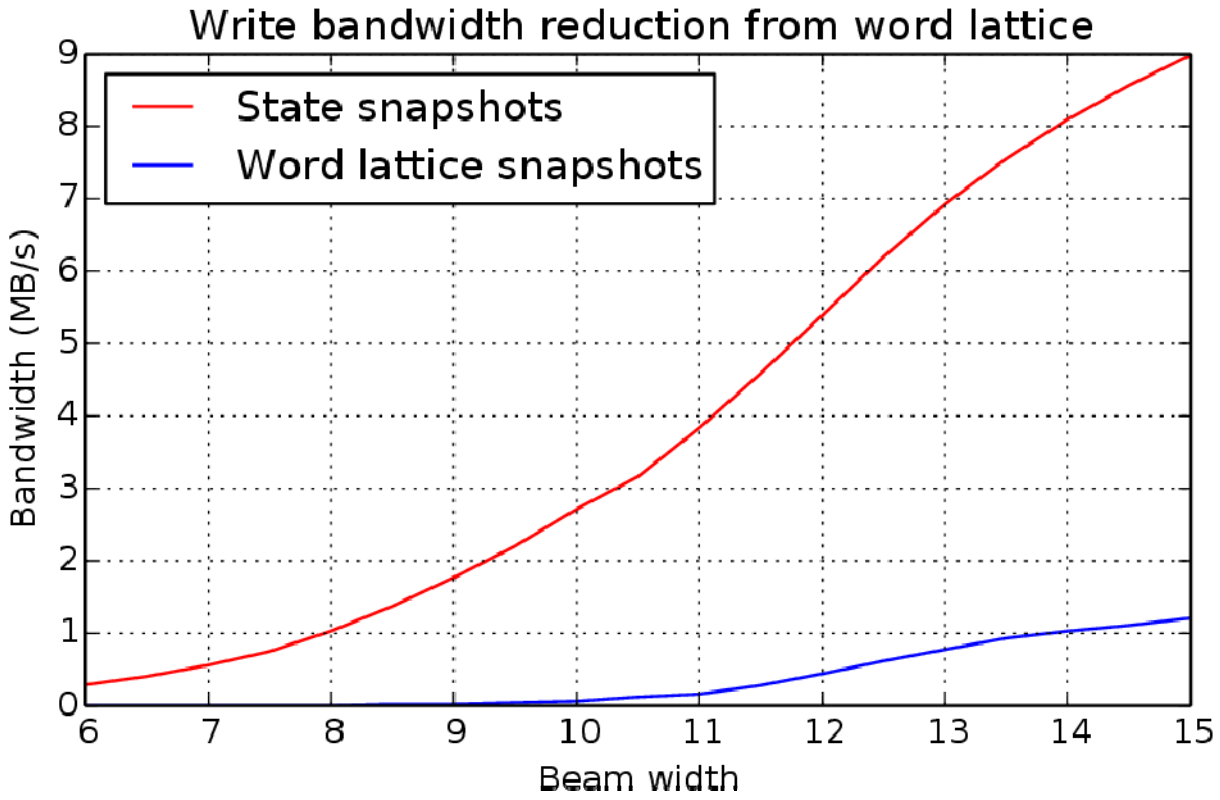


… to word lattice



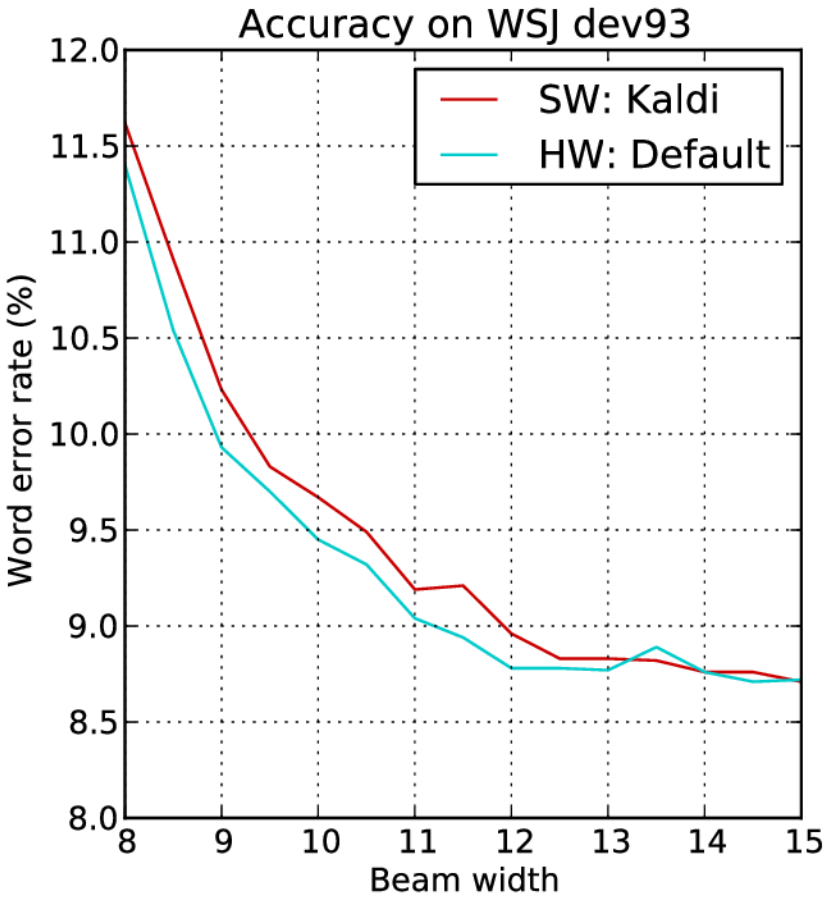
Discard intermediate states between words

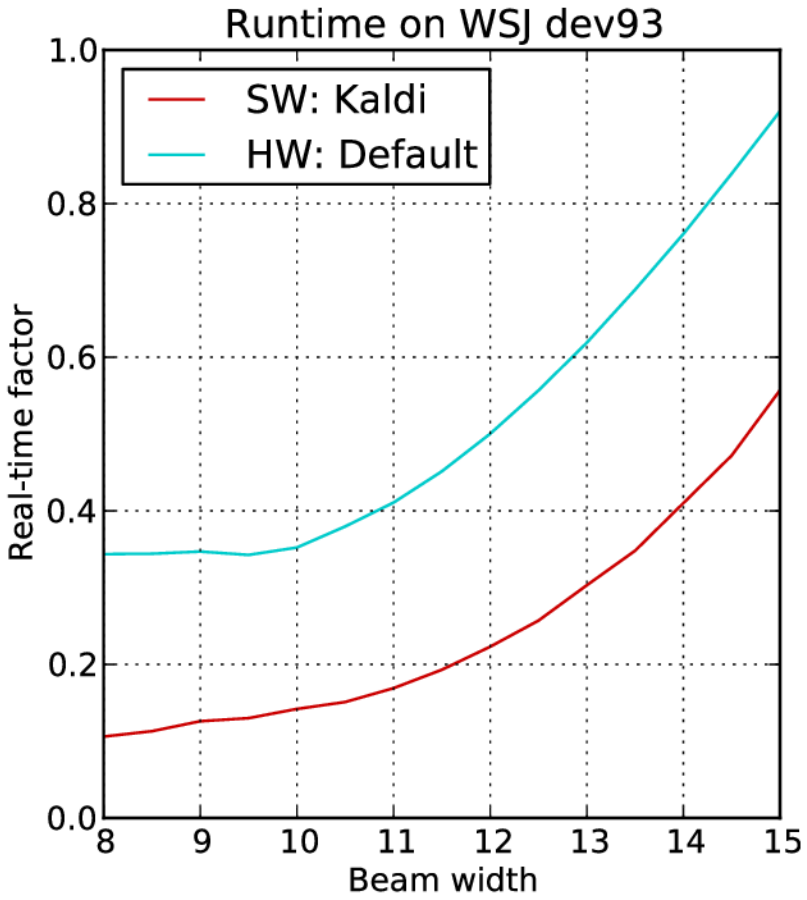
Search – Word lattice (cont.)



- **Light workloads**: no external memory writes (pruning keeps word lattice within internal memory)
- **Heavy workloads**: 8x reduction (writing only word arcs)

ASR accuracy and speed



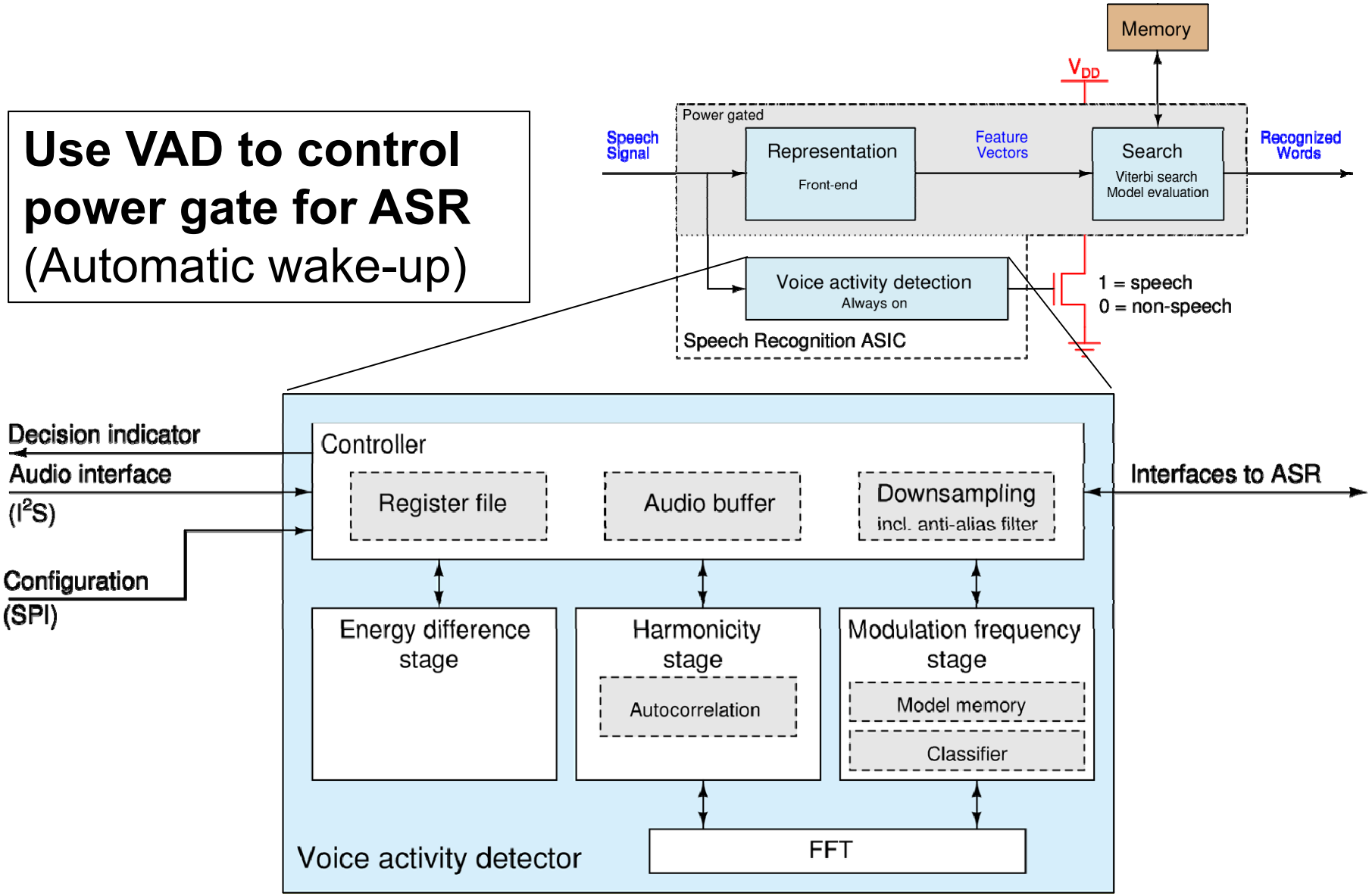


Similar accuracy to software (Kaldi) with 145k word vocab.

At 80 MHz, same search speed as software on 3.7 GHz Xeon

Voice activity detection (VAD)

Use VAD to control power gate for ASR
(Automatic wake-up)



VAD power impacts

$$P_{\text{avg}} = p_{\text{VAD}} + \left[(1 - p_M)D + p_{FA}(1 - D)\right] p_{\text{downstream}}$$

Downstream system contribution

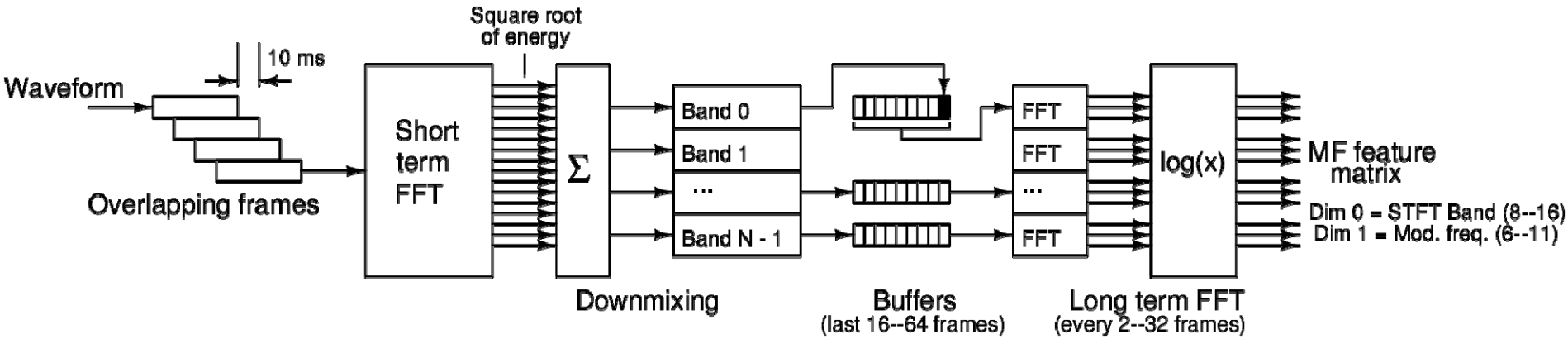
Consider typical values:

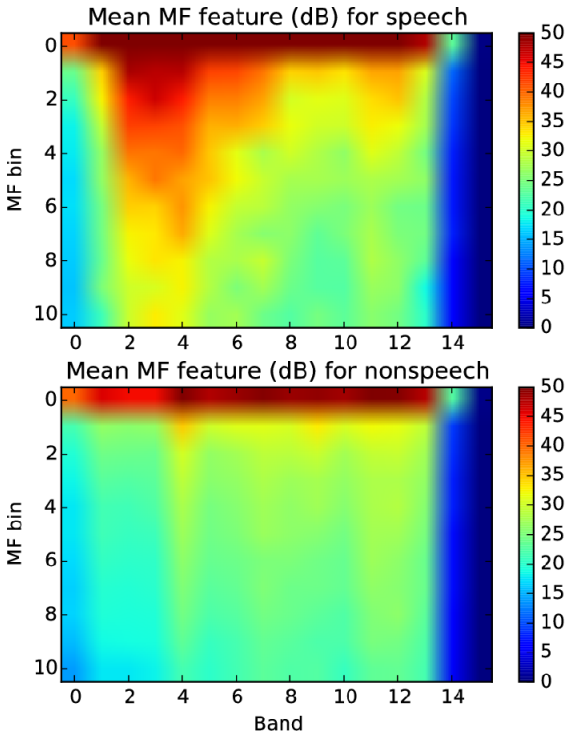
- p_{VAD} < 100 μW
- $p_{downstream}$ > 1 mW
- D < 0.05

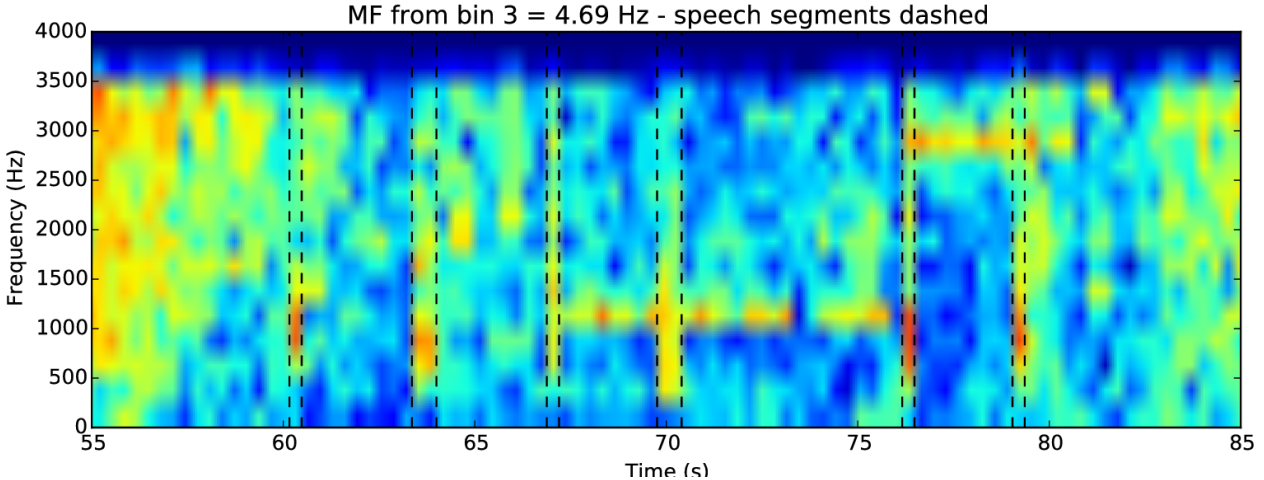
If p_{FA} is significant, averaged downstream power exceeds VAD power

- Optimize for p_M and p_{FA} (VAD accuracy) rather than VAD power

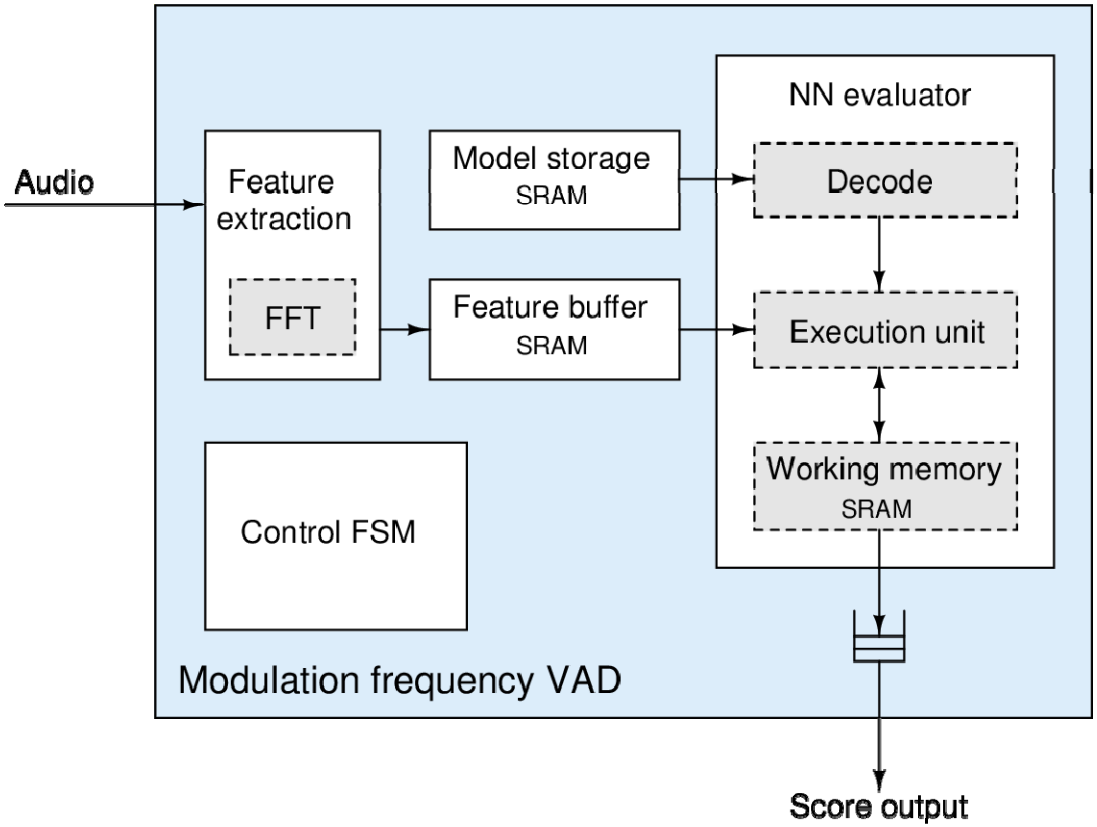
Modulation frequency VAD





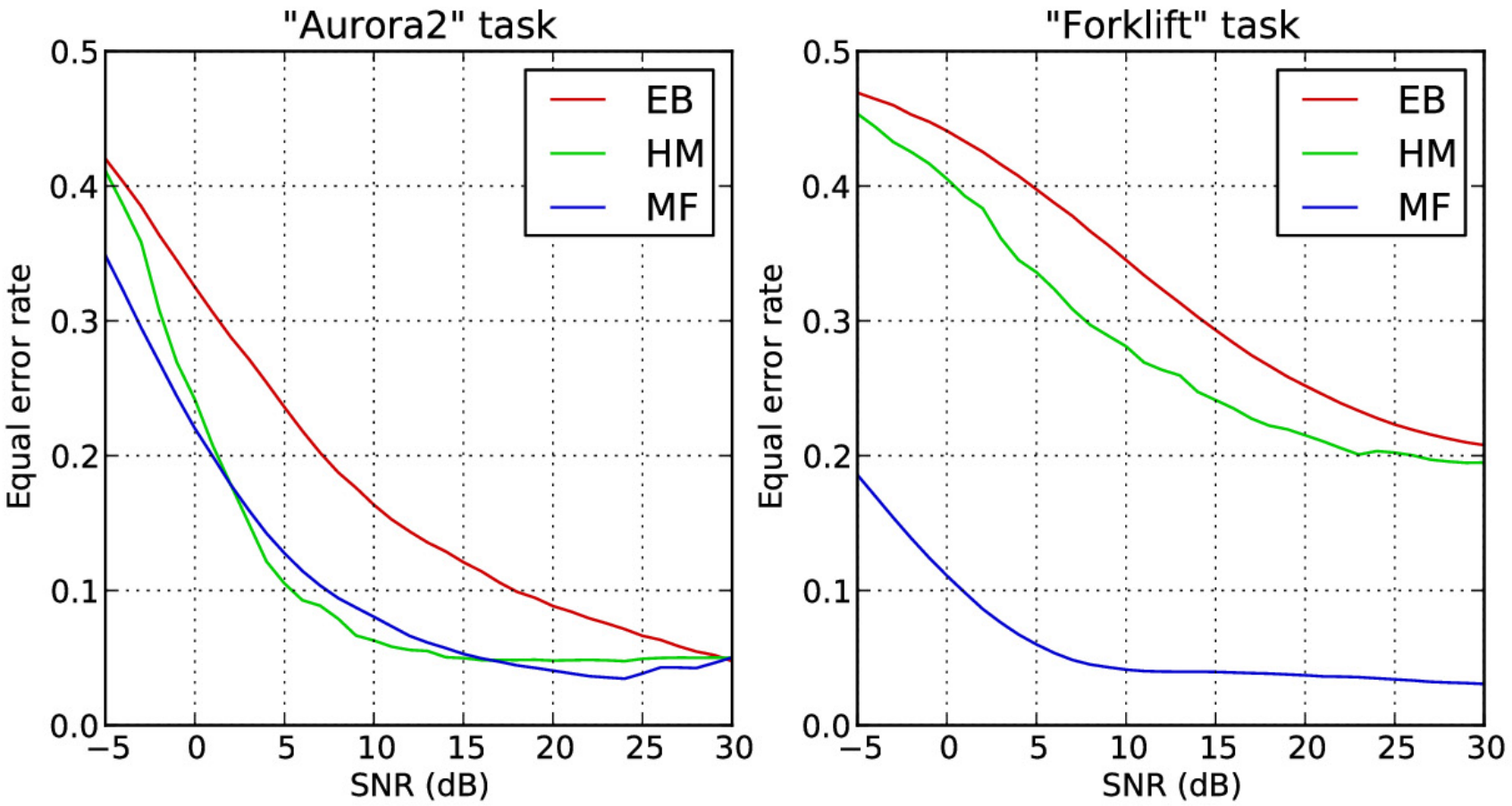


Modulation frequency VAD (cont.)



- **MF features fed to NN classifier**
  - Small network (e.g. 3x32) is sufficient
  - Fewer parameters than SVM
- **NN architecture stripped down from ASR**
  - No sparse matrix support
  - No parallel evaluation
  - Still quantized
- **No external memory**
  - Parameters supplied at startup over SPI bus

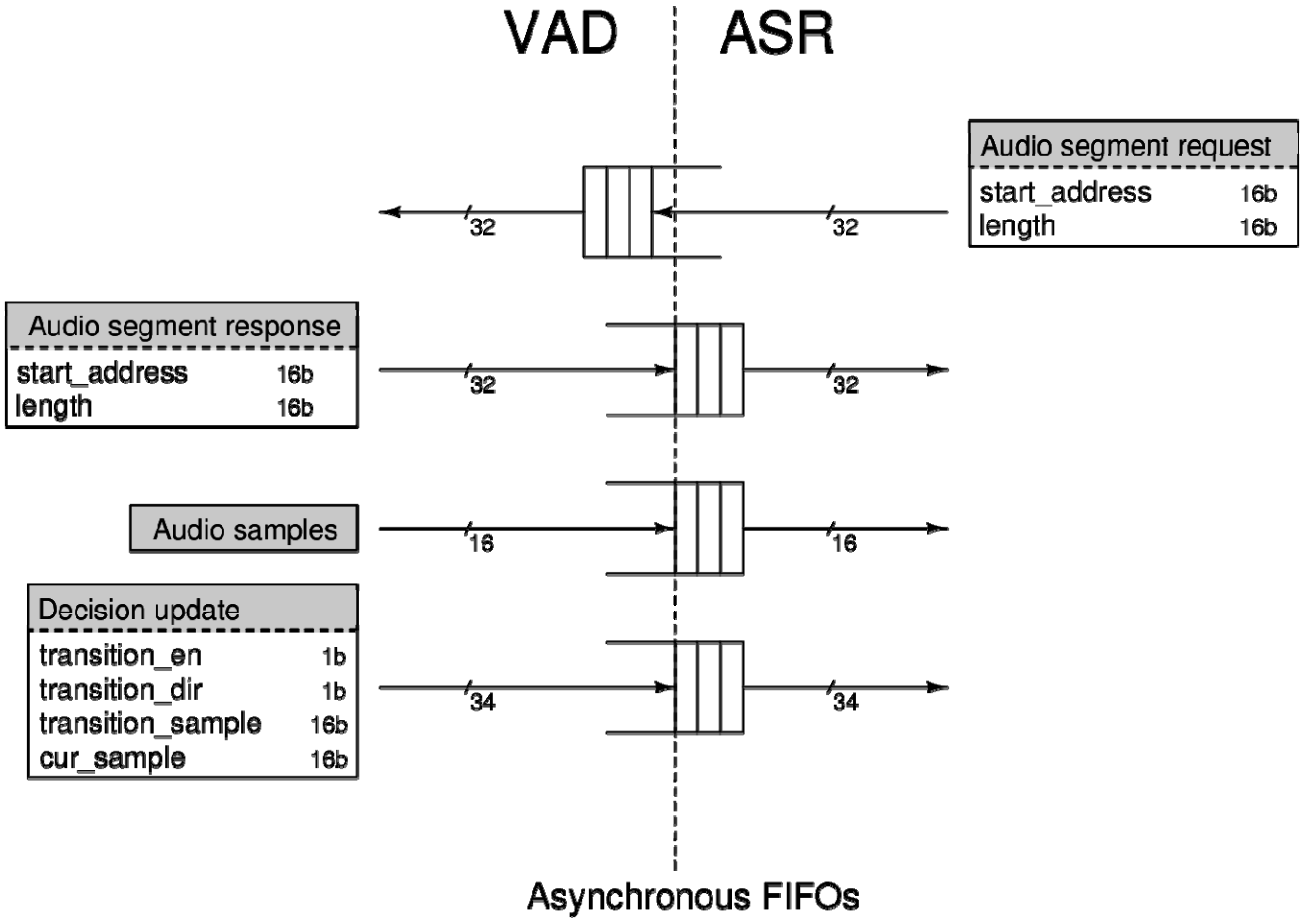
VAD performance comparison



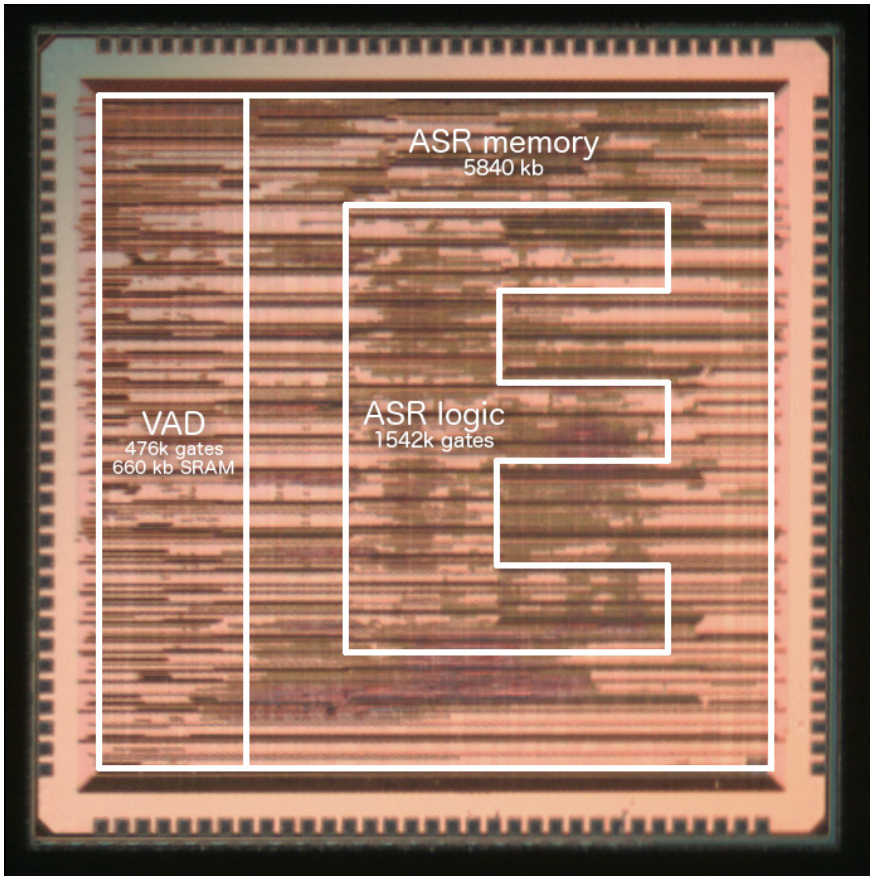
- Two tasks: Aurora2 (left), Forklift/Switchboard (right—more difficult)
- MF algorithm outperforms energy-based (EB) and harmonicity (HM)
- Performance improves with more training data

ASR/VAD interfaces

- VAD runs continuously
- Supervisory logic places ASR in reset during non-speech input
- ASR requests audio samples buffered by VAD after wake-up

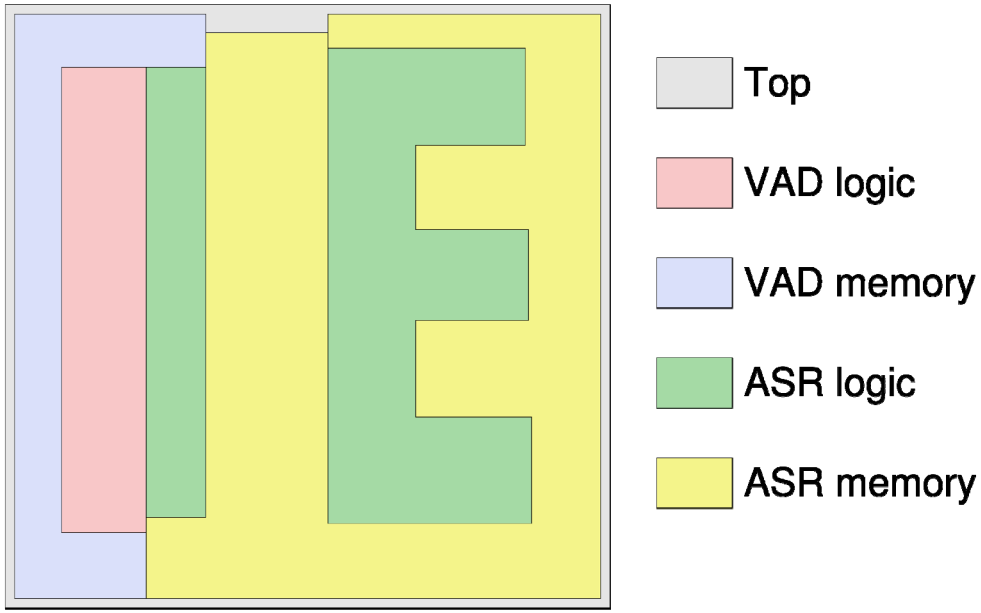


Test chip specifications

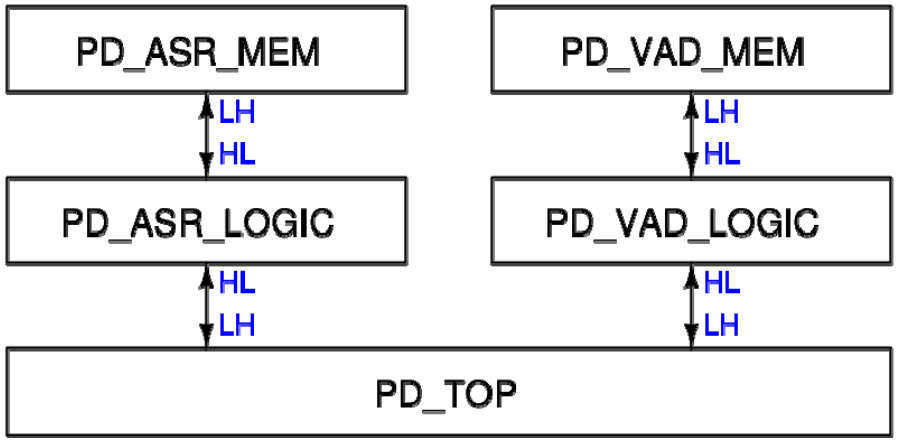


Specification	Value
Process	65 nm LP
Core size	3.1 × 3.1 mm
Die size	3.63 × 3.63 mm
Package	88-pin QFN
Logic gates	2088k (NAND2 equiv.)
SRAM	5.84 Mb
Supply voltage	0.60–1.20 V
Power consumption	1.8–7.8 mW (typ.)
Clock frequency	3–86 MHz
Neural network efficiency	16–56 pJ/neuron
Viterbi search efficiency	2.5–6.3 nJ/hypothesis

Multivoltage design



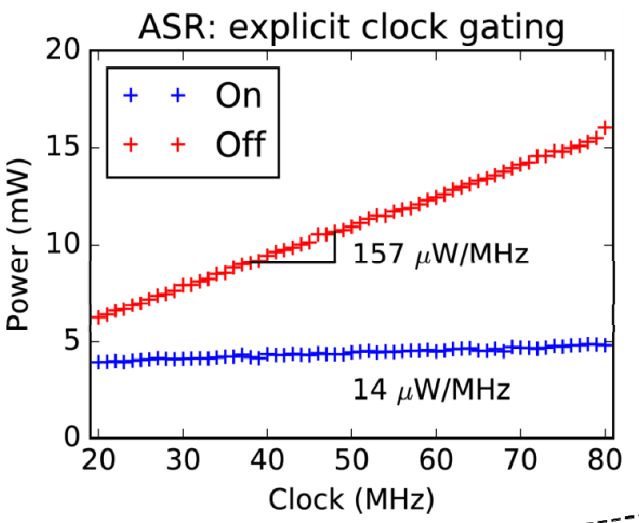
- **Memory tends to require higher voltage than logic**
  - Separate memory supplies
- **I/O level shifters (1.8—2.5 V) cannot handle low logic level**
  - Intermediate supply for top level (with supervisory logic)
- **Relationship between supply voltages is constrained**
  - Level shifters operate in one direction

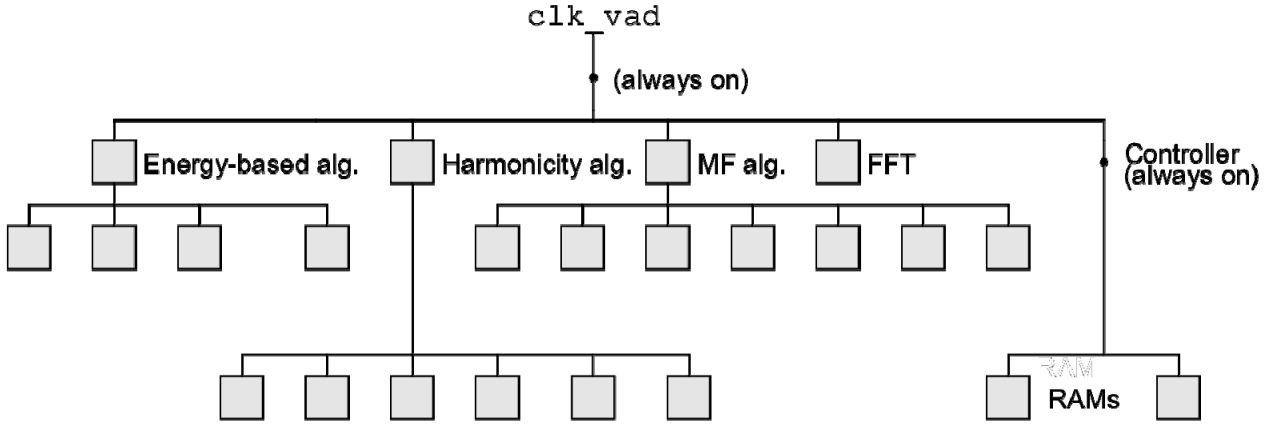


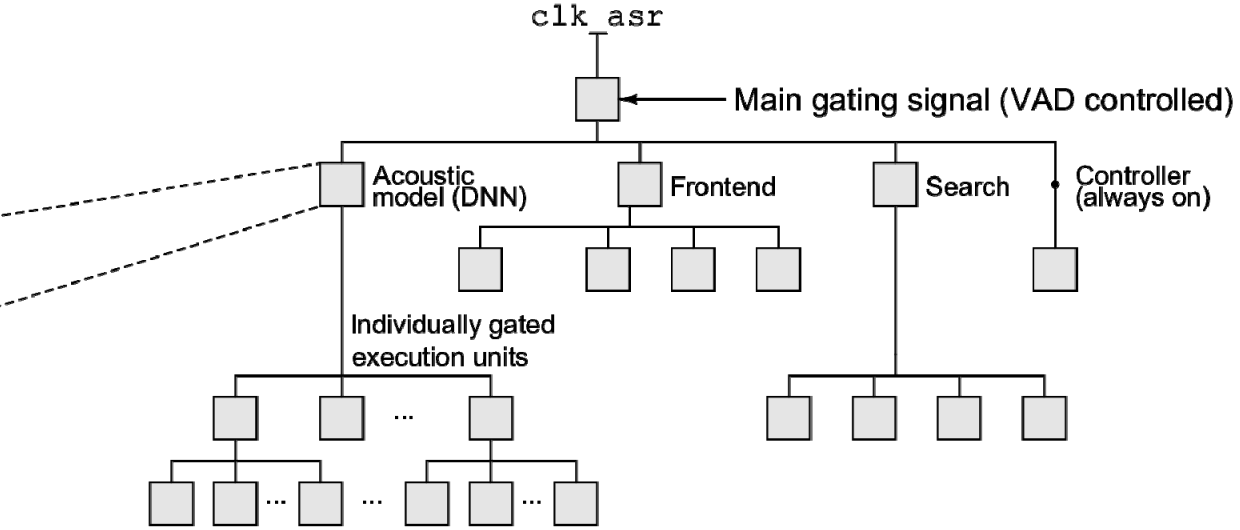
Level shifters

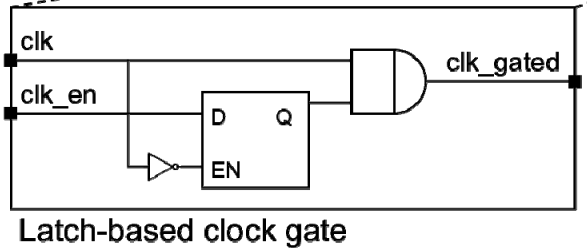
LH: low to high
HL: high to low

Clock gating



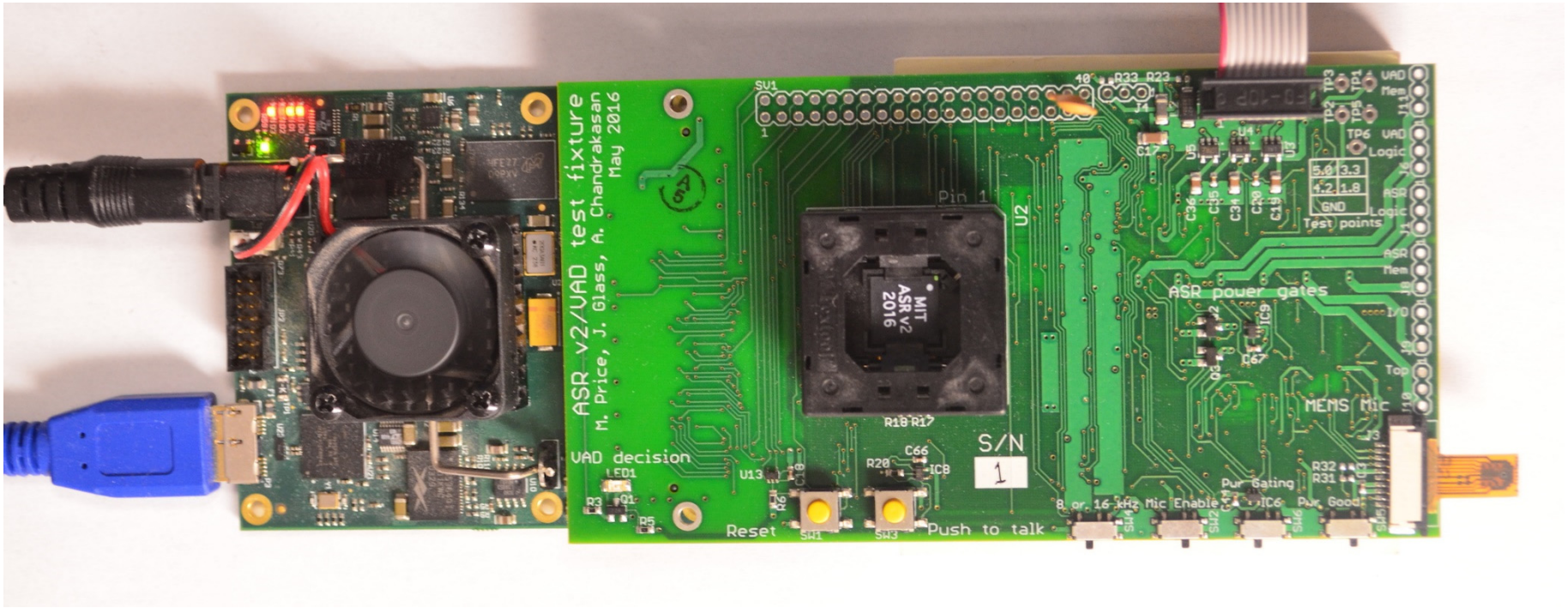






Explicit clock gating complements automatic clock gating

Test setup



- On-board control of clocks and power supplies
- FPGA provides host and memory interfaces
- Tested live audio, real-time streaming, batch processing

Measured performance and scalability

Note: ASR model sizes and search parameters vary between tasks; ASR and VAD tested independently

Automatic speech recognition (ASR)

Task	Vocab.	Clock	Mem. BW	WER	Power
Digits	11	3 MHz	0.11 MB/s	1.65%	172 μW
Weather	2k	23 MHz	10.1 MB/s	4.38%	4.70 mW
Food diary	7k	46 MHz	9.02 MB/s	8.57%	4.67 mW
News (1)	5k	15 MHz	4.84 MB/s	3.12%	1.78 mW
News (2)	145k	40 MHz	15.0 MB/s	8.78%	7.78 mW

45x range

Voice activity detection (VAD)

	SNR for 10% frame level EER			Power
Algorithm	White noise	Aurora2	Forklift	
Energy-based	-1 dB	18 dB	Fail	8.5 μW
Harmonicity	< -5 dB	5 dB	Fail	24.4 μW
Modulation frequencies	-2 dB	7 dB	1 dB	22.3 μW

Conclusions

- **DNNs can improve performance of HW ASR**
  - Even with restricted memory bandwidth: < 10 MB/s
- **DNN based VAD can be robust and compact**
  - Model stored on-chip (< 12 kB)
  - Low power (22.3 μW)
- **ASR is not only about neural networks**
  - Significant effort required for feature extraction and search
  - NN architecture developed with knowledge of application
- **Combination of algorithm, architecture, and circuit techniques delivers:**
  - Improved accuracy – fewer word errors
  - Improved programmability – train using standard tools
  - Improved scalability – lower power consumption