NeoNexus: The Next-generation Information Processing System across Digital and Neuromorphic Computing Domains

Qinru Qiu
Dept. of Electrical Engineering and Computer Science, Syracuse University

Hai Li, Yiran Chen
Dept. of Electrical and Computer Engineering, University of Pittsburgh

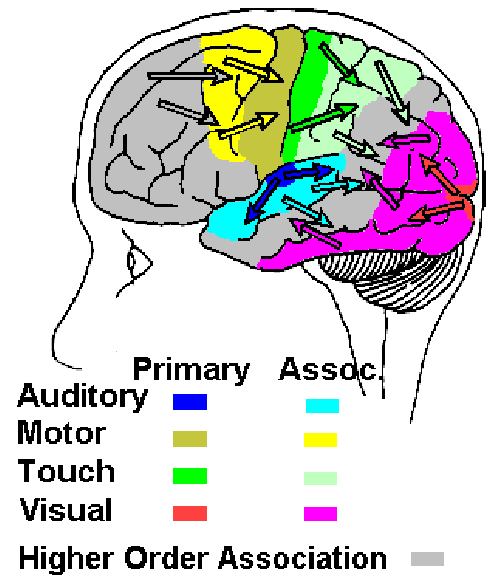
Brain Inspired Computing

- The performance of traditional Von Neumann machine is reaching to a limit
- Human neocortex system has unprecedented performance and power efficiency
  - Particularly in language understanding, image recognition and situation awareness

$\neq$

Brain Inspired Information Processing

- Brain inspired information processing relies on two main operators
  - Pattern detection
  - Probabilistic inference
- Multiple stages in human sensory processing
  - Primary sensory cortex detects a specific input (i.e. contour, color, or pitch, etc.)
  - Association cortex combines information from primary sensory cortex to produce perception
  - Higher order association combines different sensory association areas



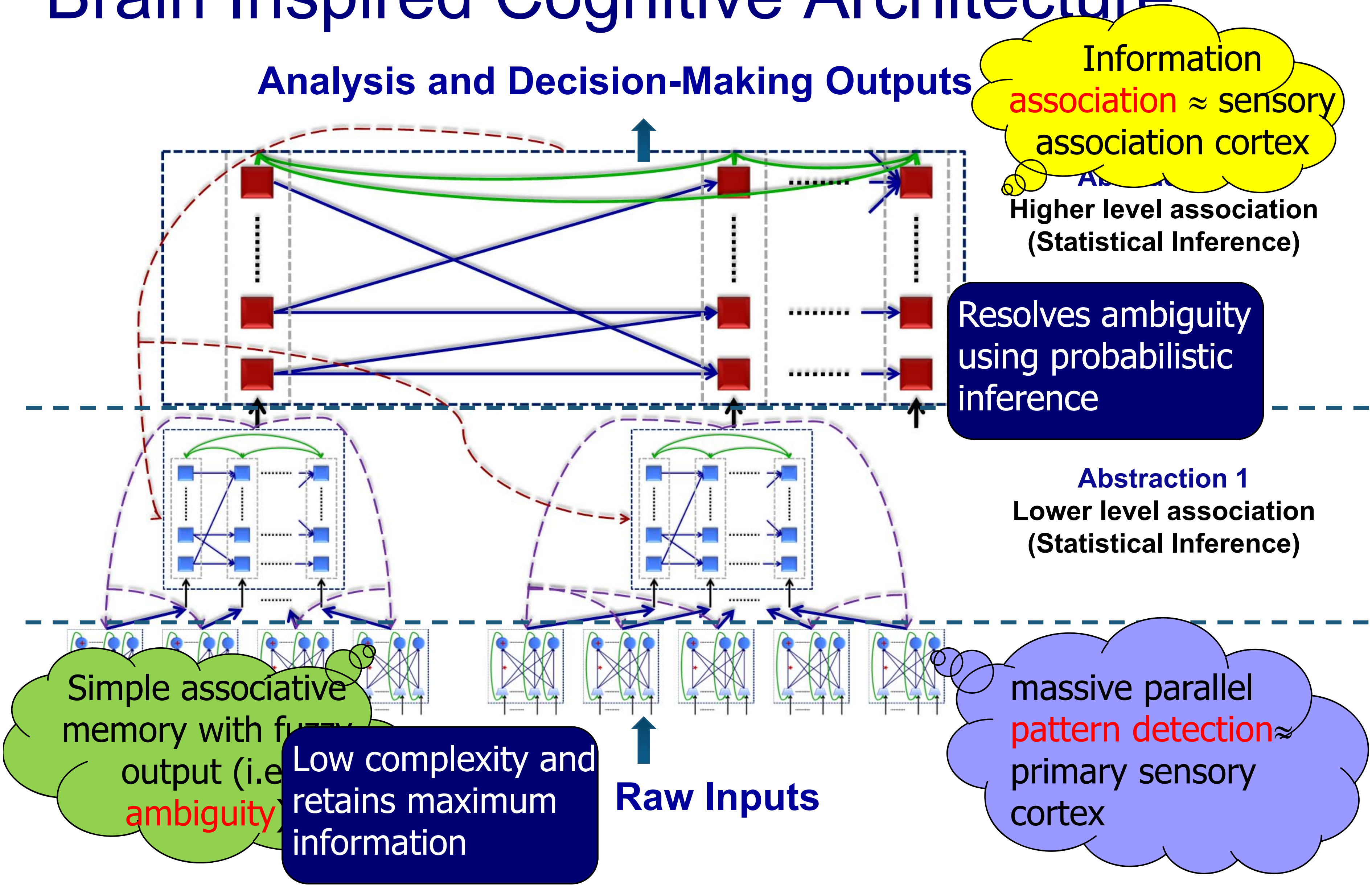
knowledge

Key Features of Neuromorphic Computing

- Performs pattern detection and probabilistic inference
- Massive parallel
- Closely coupled storage and computation
- Distributed storage with high redundancy provides reliability
- Simple unified building blocks (i.e., neurons)
- Analog/mixed signal domain operation

We need non-conventional solutions for both hardware architectures and software computation models

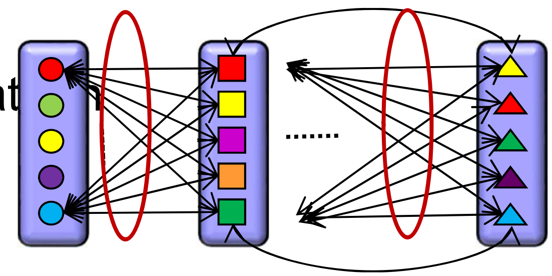
Brain Inspired Cognitive Architecture



Computation Models

Matrix-vector multiplication

Knowledge matrices KL



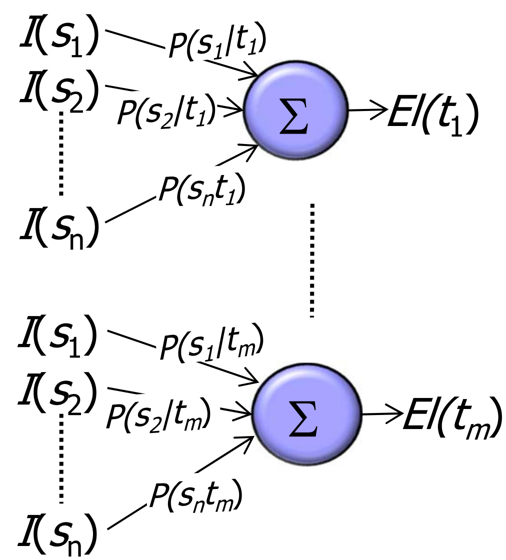
- Bottom layer: BSB (Brain-State-in-a-B
  - Convergence speed gives fuzzy information about pat similarity
- Upper layer: probabilistic inference
  - Features and attributes represented as lexicons and symbols
  - Association among features represented by knowledge links
    - Captures $\log[p(s_i \mid t_j)]$ between source and target symbols

Redundancy in KL provides reliability

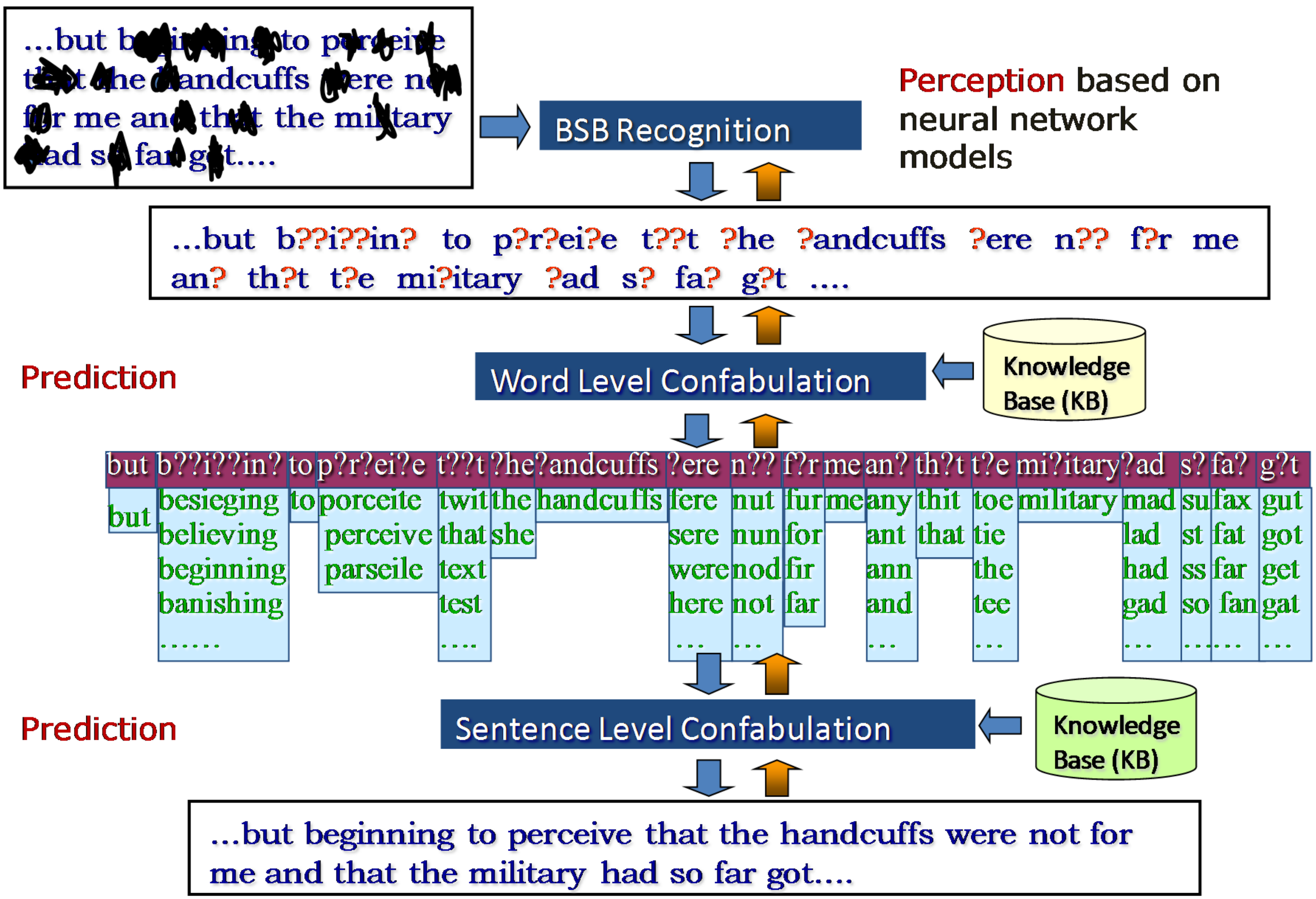
- Analogies to neocortical system
  - Symbols ⇔ neurons
  - Knowledge link ⇔ synapses
  - Knowledge link values ⇔ Hebbian plasticity
  - Symbols in same lexicon ⇔ neurons with inhibition link
  - Symbols in lexicon ⇔ neurons with excitation link
  - Likelihood ca and belief prop Integration-and-fire with soft-winner-takes-all

Matrix-vector multiplication

Comparison and sorting

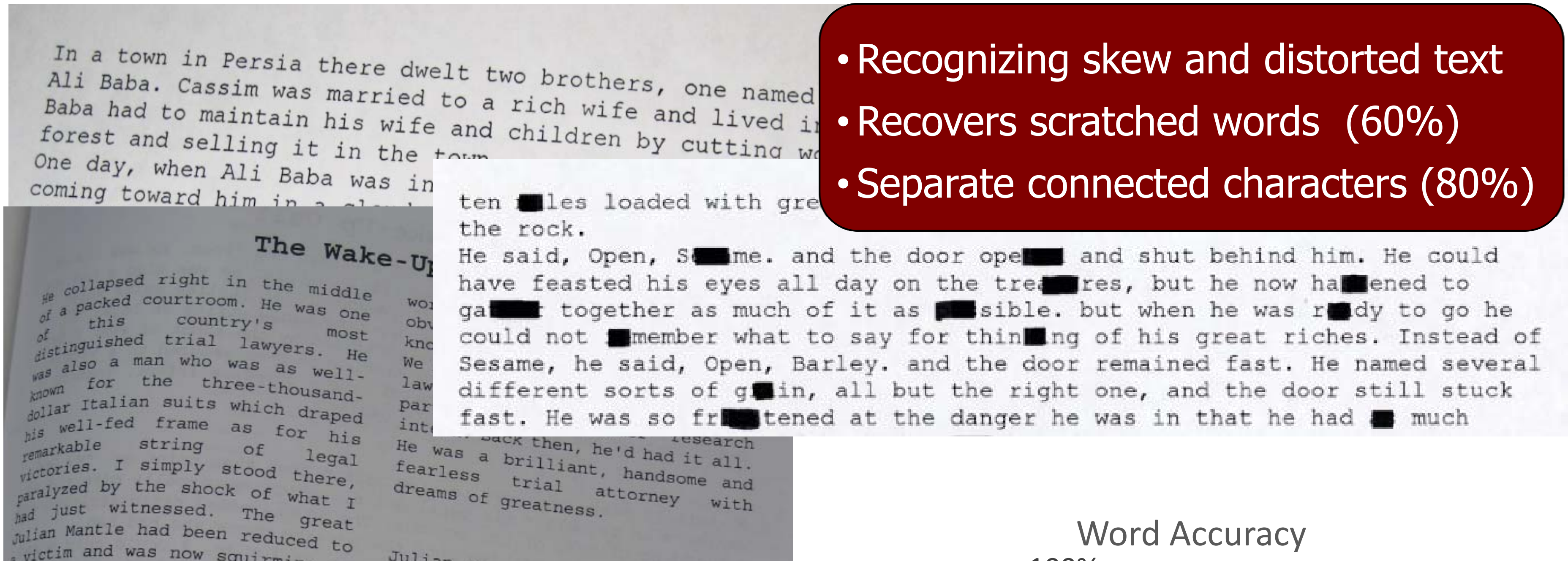


Context Aware Intelligent Text Recognition



Recall Accuracy

- Recognizing skew and distorted text
- Recovers scratched words (60%)
- Separate connected characters (80%)



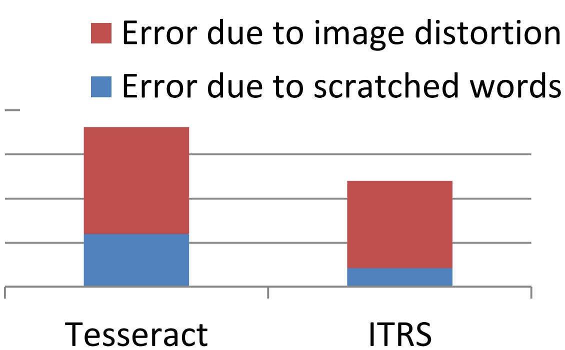
In a town in Persia there dwelt two brothers, one named
Ali Baba. Cassim was married to a rich wife and lived i
Baba had to maintain his wife and children by cutting w
forest and selling it in the town
One day, when Ali Baba was in
coming toward him in a cloud

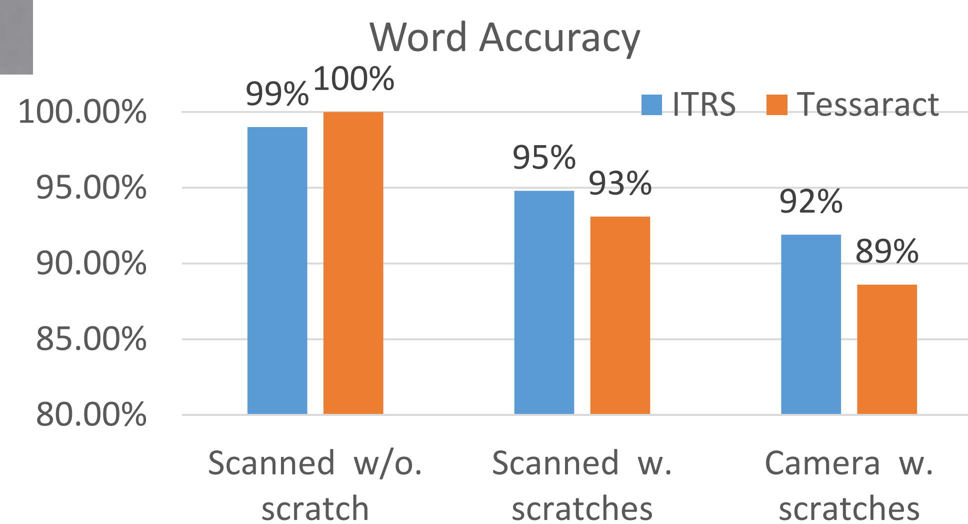
The Wake-Up

He collapsed right in the middle of a packed courtroom. He was one of this country's most distinguished trial lawyers. He was also a man who was well-known for the three-thousand-dollar Italian suits which draped his well-fed frame as for his remarkable string of legal victories. I simply stood there, paralyzed by the shock of what I had just witnessed. The great Julian Mantle had been reduced to a victim and was now squirming

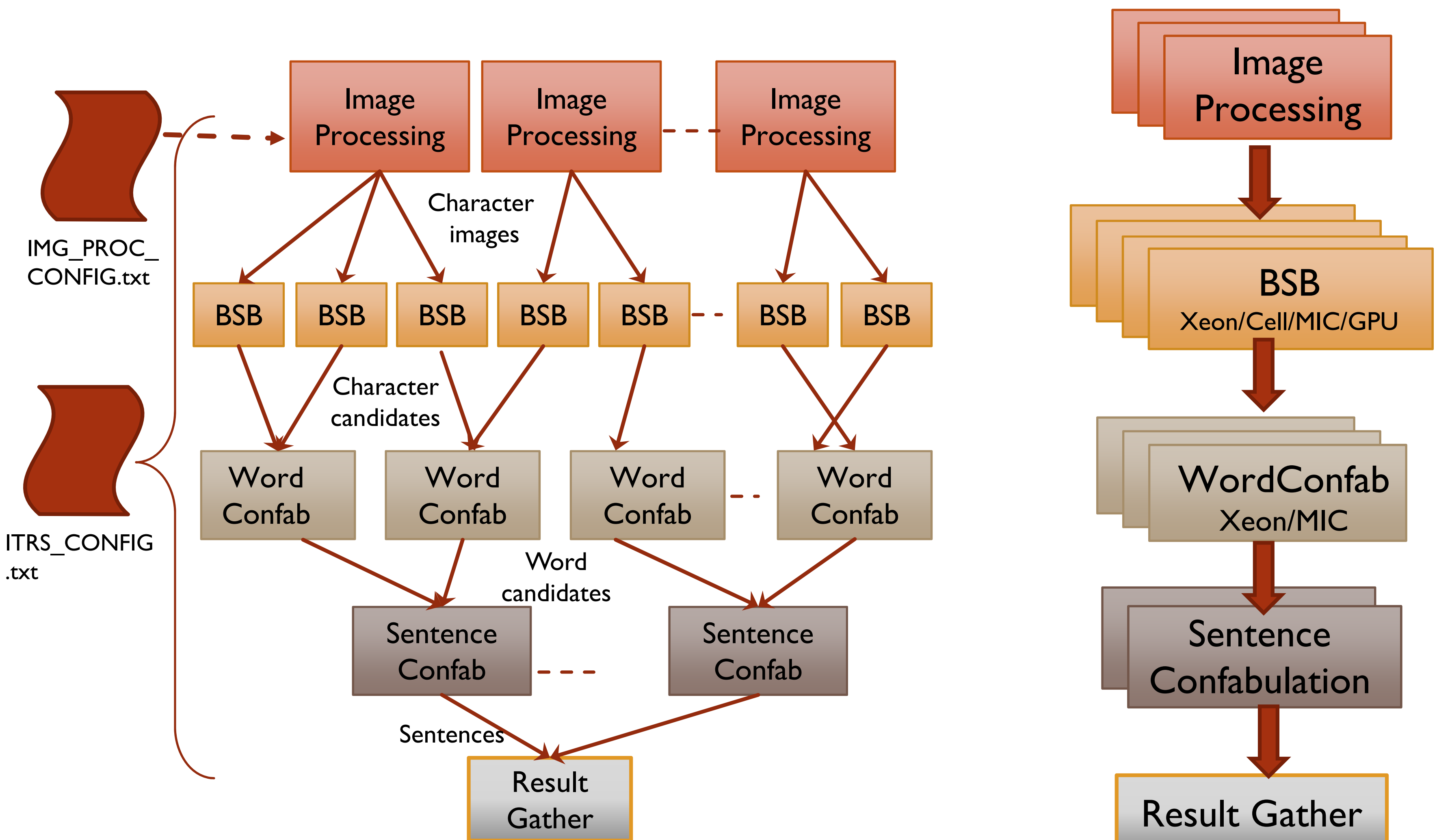
back then, he'd had it all. He was a brilliant, handsome and fearless trial attorney with dreams of greatness.

ten [redacted]les loaded with gre
the rock.
He said, Open, S[redacted]me. and the door ope[redacted] and shut behind him. He could
have feasted his eyes all day on the trea[redacted]res, but he now ha[redacted]ened to
ga[redacted]t together as much of it as [redacted]sible. but when he was r[redacted]dy to go he
could not [redacted]member what to say for thin[redacted]ng of his great riches. Instead of
Sesame, he said, Open, Barley. and the door remained fast. He named several
different sorts of g[redacted]in, all but the right one, and the door still stuck
fast. He was so fr[redacted]tened at the danger he was in that he had [redacted] much

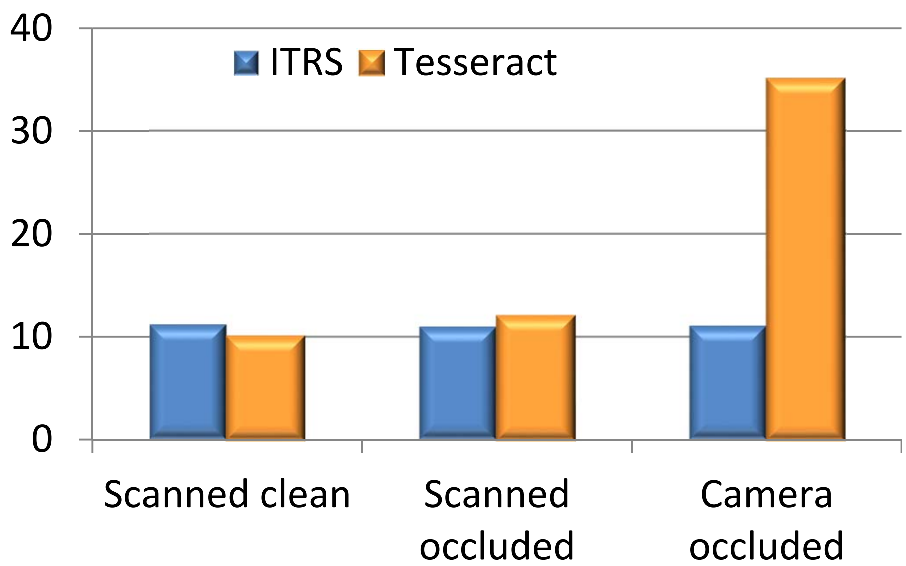




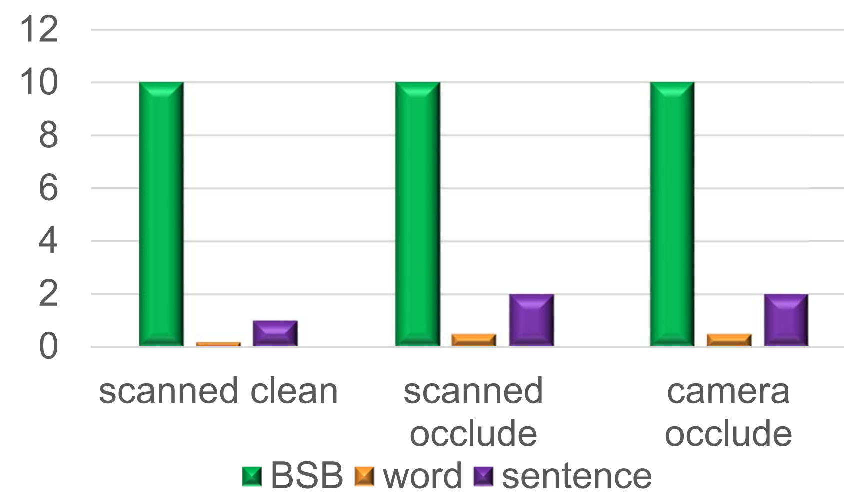
On Multicore Heterogeneous Architecture



Processing Time

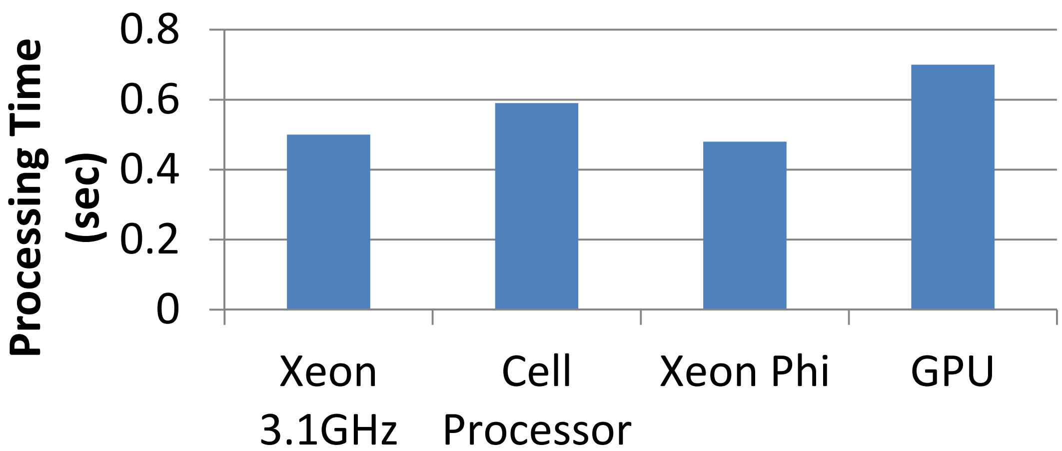


- Configuration: BSB (MIC0), Word (MIC1), Sentence (Xeon)
- The processing time of Tesseract rises rapidly as the image size increases and image quality reduces
- The processing time of ITRS remains stable



- Sentence and word confab time increases as the image quality reduces
- BSB processing is the bottleneck in ITRS

Performance Comparison



1 workload = checking 96 images against 93 patterns ≈ 58×10^9 floating point operations

	Xeon	Cell	Phi	GPGPU
Clock Frequency (GHz)	3.1	3.2	1.1	0.575
Number of Physical Cores	8	7	61	14
Number of Logical Cores	32	7	244	448
Peak Performance (TFLOPS)	~0.5	~0.2	~2	~1.0
Sustained Performance (GFLOPS)	116	96	128	83
Utilization	23%	48%	6.4%	8.3%

Brain-inspired Anomaly Detection

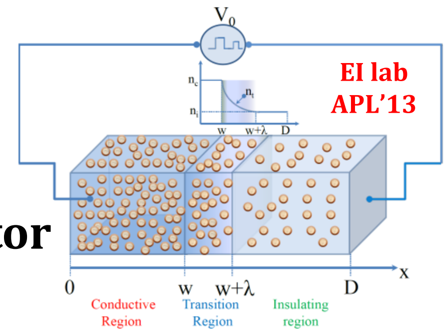
- An anomaly is a surprise
  - Something different from expectation
    - An attribute with low likelihood
- Likelihood-ratio test for anomaly detection
  - x is abnormal if it is less likely to be observed than a_i, $\exists a_i \in A$
    - x: observed attribute, A: the set of all potential attributes
  - Anomaly score: $$\frac{\max_i[el(ai)] - el(x)}{\max_i[el(ai)]}$$
  - A high anomaly score means relatively less likely event
  - Successfully applied to vehicle monitoring and cyber security detection

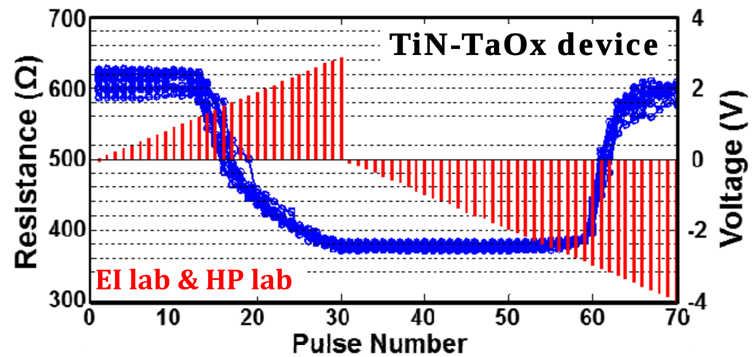
Observations

- Matrix-vector multiplication is the dominant operation in both layers:
  - Pattern matching layer: dense matrix, dense vector, consistent in matrix size
  - Inference layer: sparse matrix, sparse vector, large variations in size
- No intra-layer communication within pattern matching layer
- Frequent intra-layer communication is needed in association layer for belief propagation/likelihood estimation
  - Delay insensitive
  - Lexicons can work asynchronously
- Computation complexity of inference layer reduces as more features are considered
  - Example:
    - Sentence completion based on only language features requires at least 12-bit fix-point representation of knowledge value
    - Sentence reconstruction in ITRS, binary representation of knowledge value gives good results
  - Use additional knowledge / sensory information to reduce computation
  - Input specific computing kernel

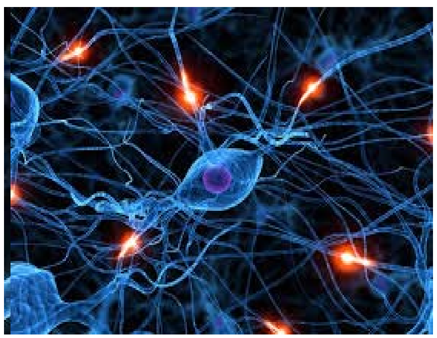
Memristor – Rebirth of Neuromorphic Circuits

Memristor

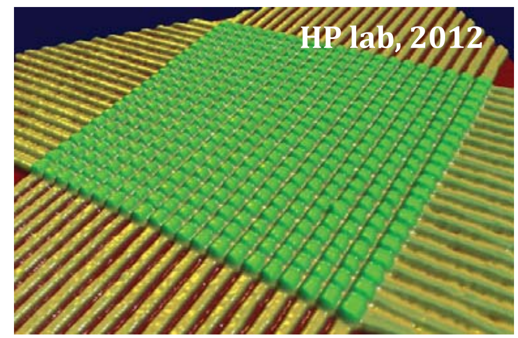




Programmable resistor w/ analog states



Synapse Network ⟺ **Memristor Crossbar**



Natural matrix operation

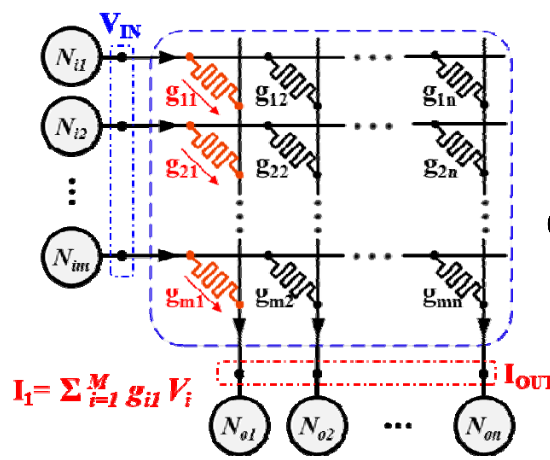
$$[x_1 \; x_2 \; \dots \; x_m] \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ g_{m1} & g_{m2} & \cdots & g_{mn} \end{bmatrix}$$

$$\|$$

$$[\; y_1 \quad y_2 \quad \dots \quad y_n \;]$$

$$y_1 = \Sigma\, x_i \cdot g_{i1}$$

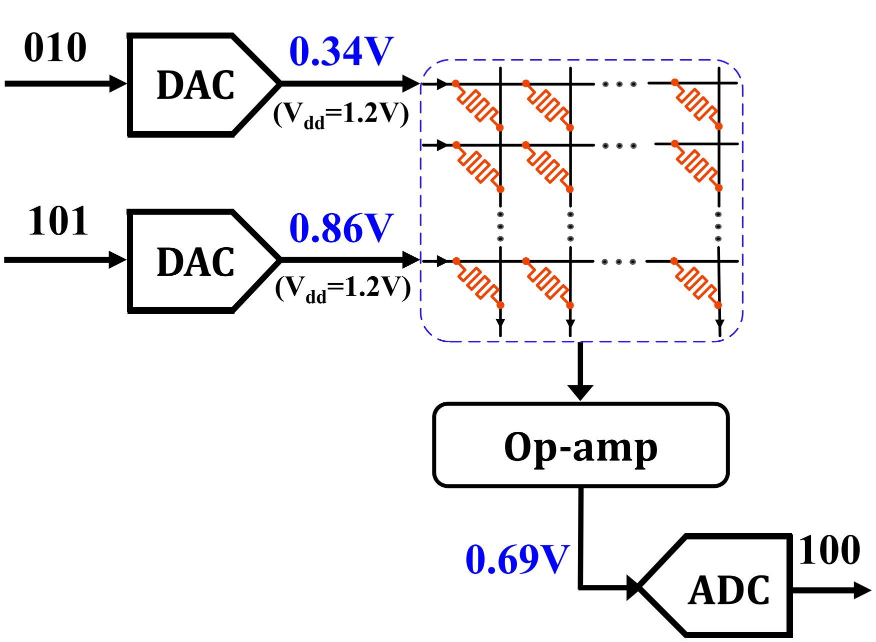
EI lab
DAC'12



High density

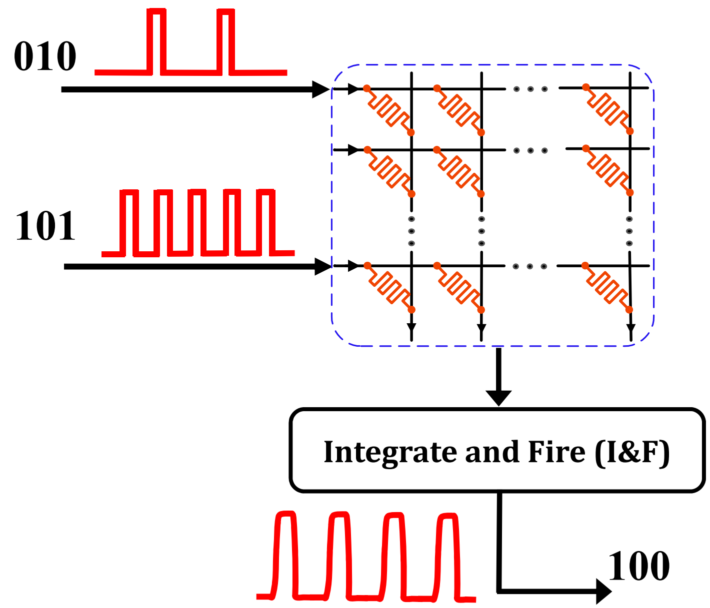
Two Design Approaches

Level-base Design



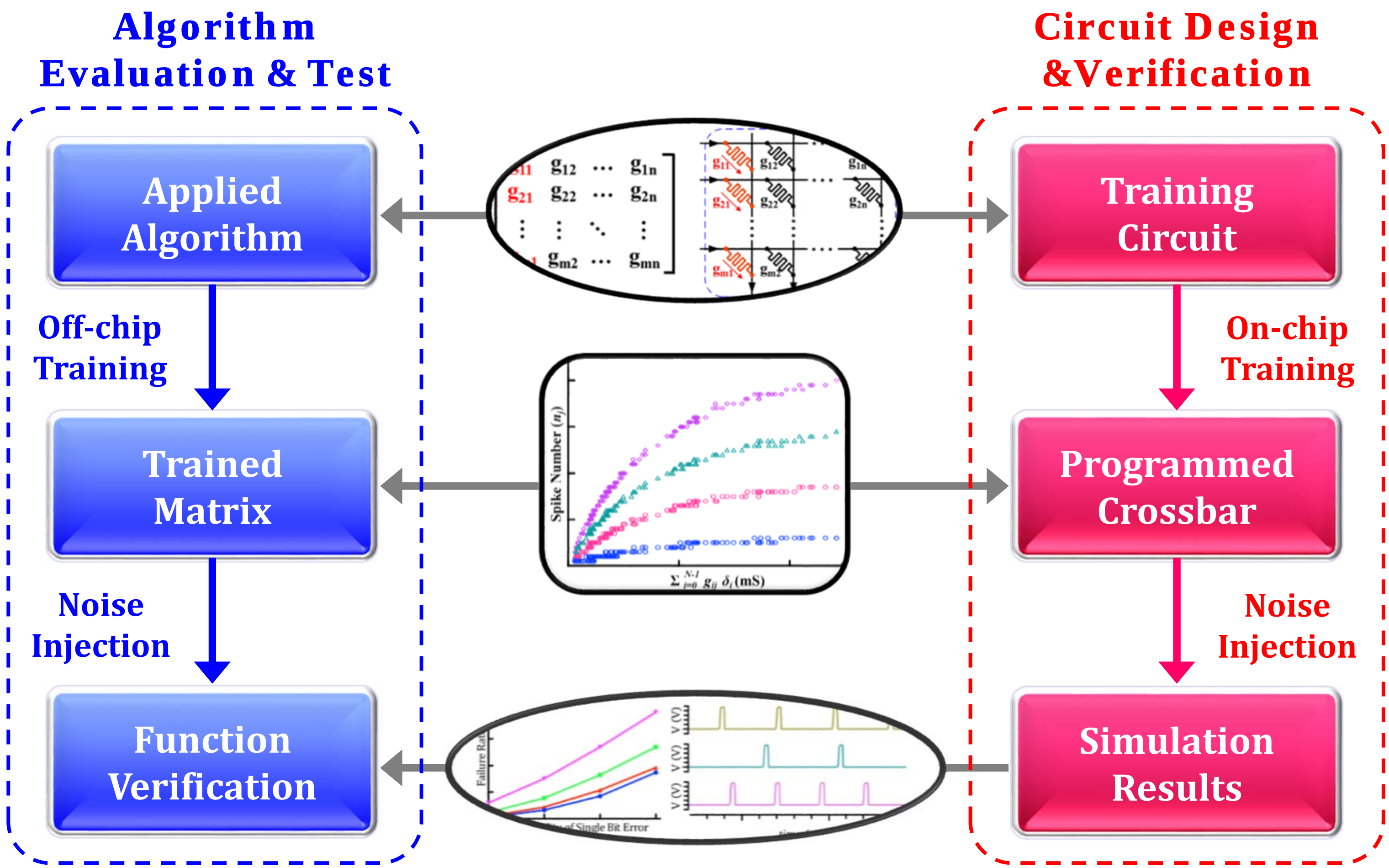
- Compatible to existing signal processing
- High speed computation

Spike-base Design



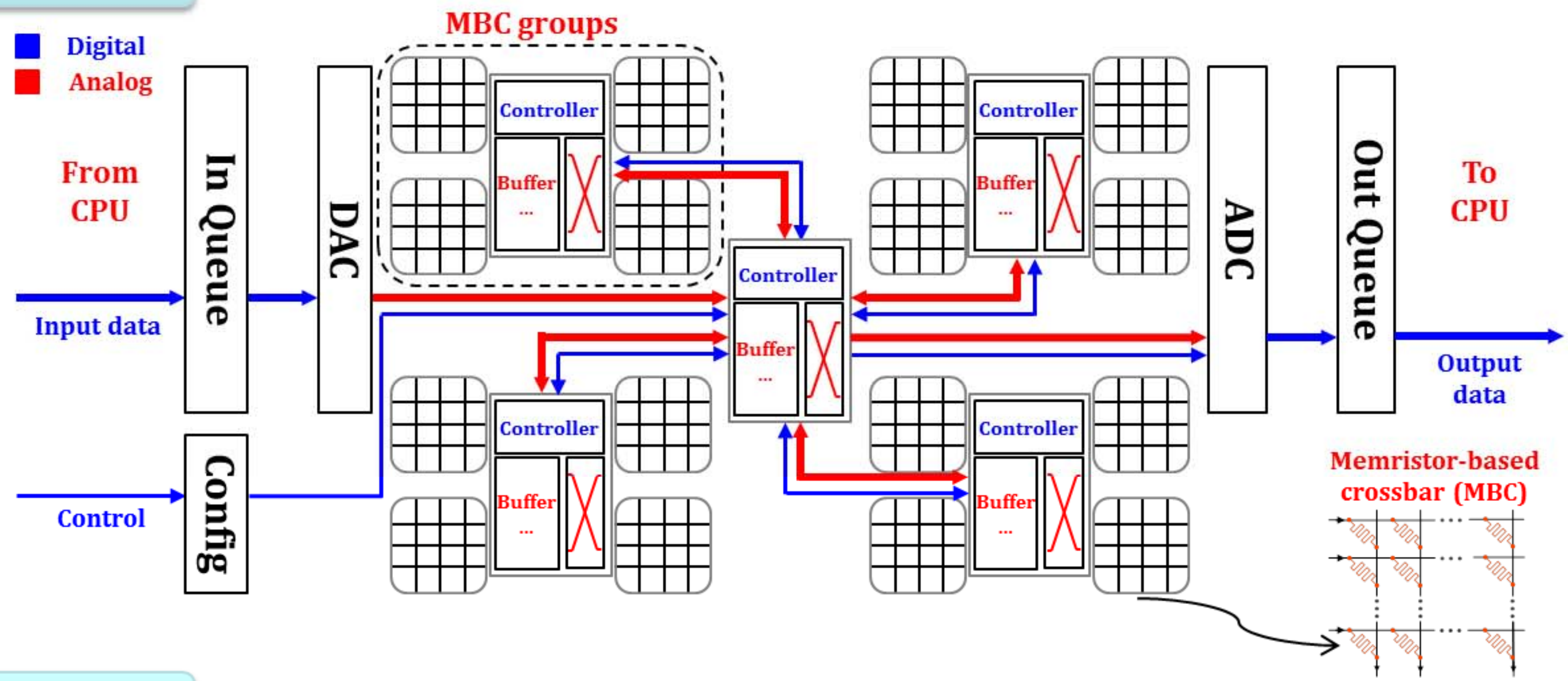
- Closer to biological system
- Extremely high power efficiency

A Cross-Optimization Design Flow



Neuromorphic Computing Acceleration (NCA)

NCA Hardware



NCA Software

```
bool Recall(float *vec, float *wm)
{   /* simulate the synapse network*/
    for(i=0;i<BsbSize;++i) wx[i] += □
wm[i*BsbSize+j] * vec[j];
    ......
```

Find the candidate codes

Source-to-source translation

```
bool Recall(float *vec)
{
   Send(NCA.id, vec);
   return Receive(NCA.id)
   ......
```

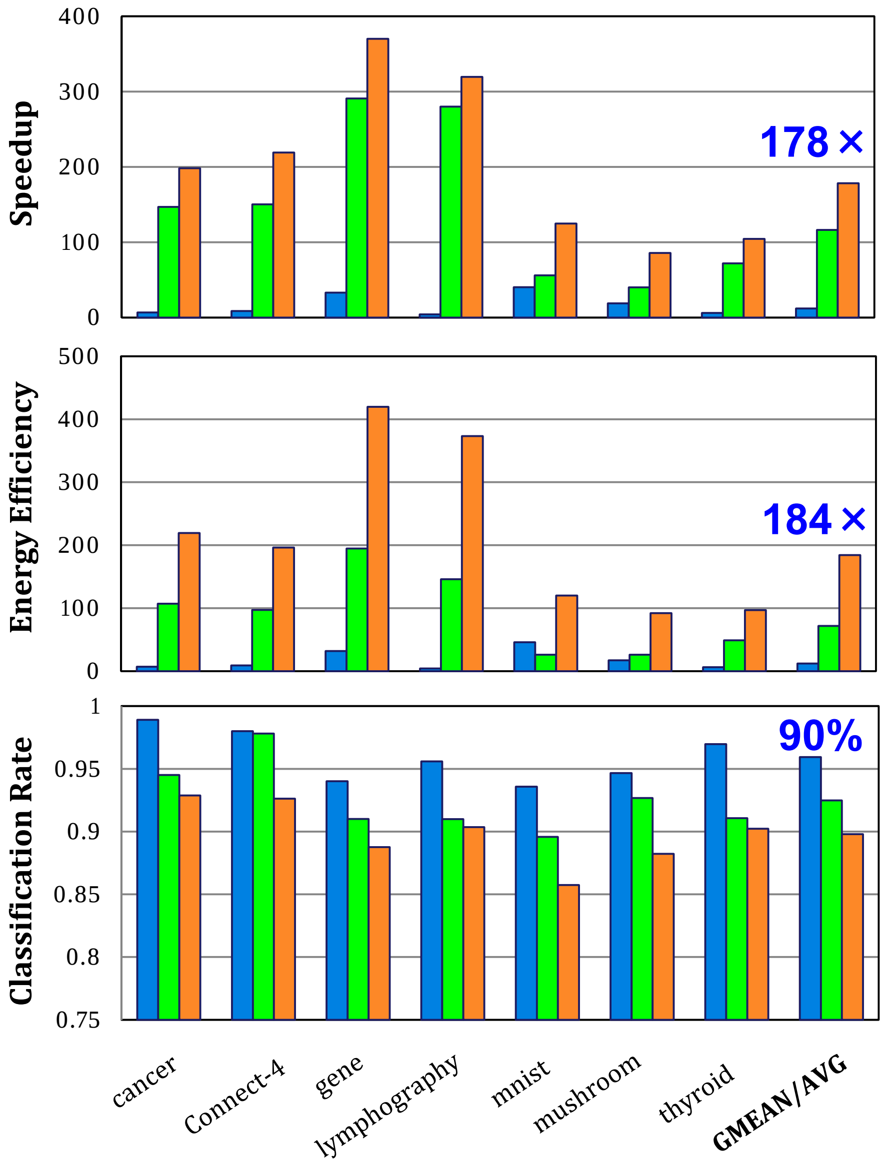
The neural topology

NCA-aware compilation

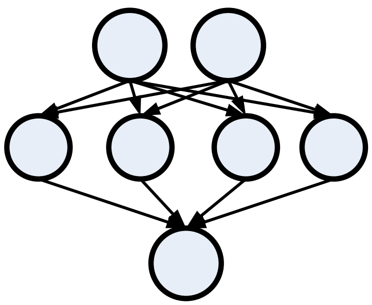
```
MOVD NCA.id, R1
......
SET NCA.id, #VAL
LAUNCH
DEQ R1, NCA.id
```

The NCA-aware executable

Compare to Other Designs



Example: Multilayer Perception (MLP)



Seven representative learning benchmarks.

All the results are normalized to the baseline CPU.

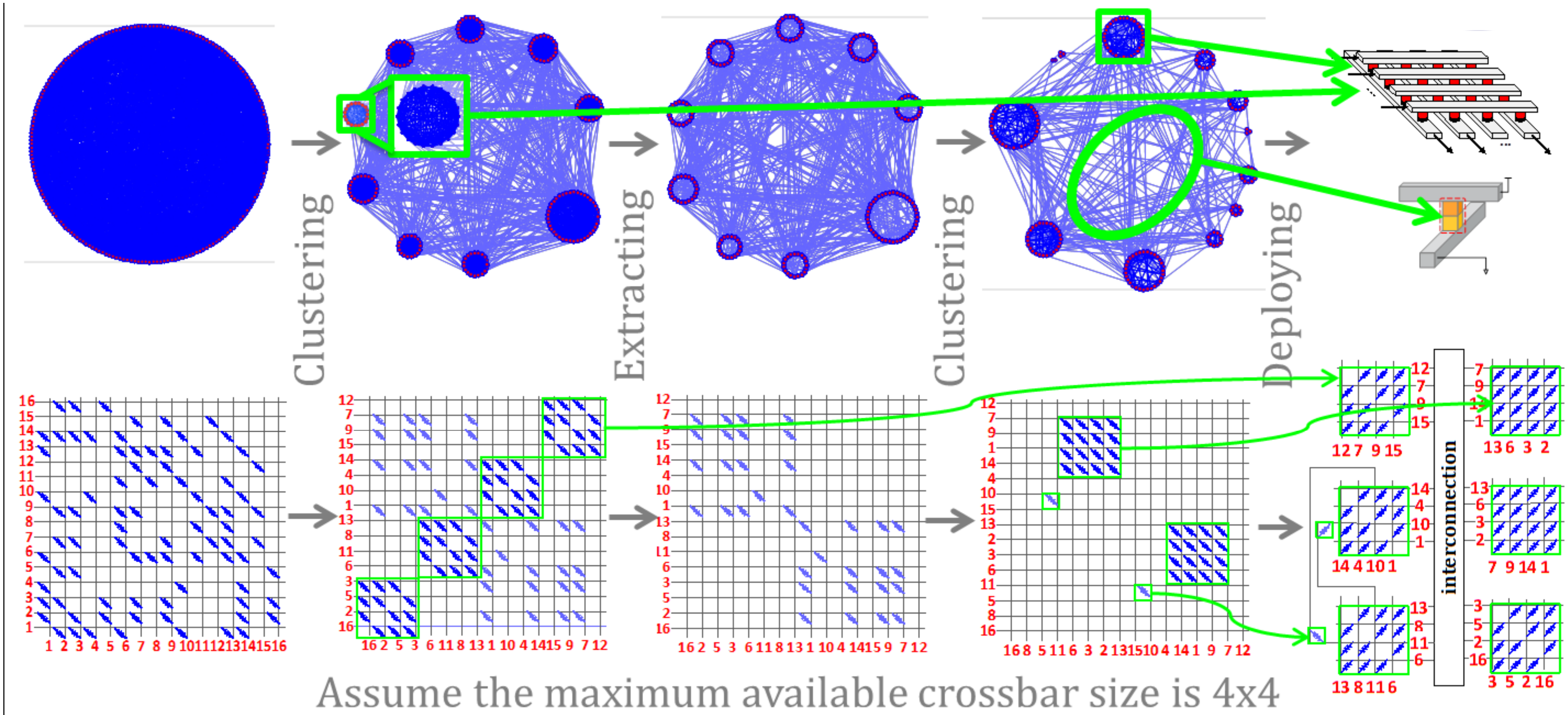
Digital NPU + Dgital NoC [1]

MBC + Digital NoC

NCA (MBC + Mixed-signal NoC)

[1] H. Esmaeilzadeh *et al.*, MICRO'12

Neuron Clustering



16 crossbars -> 6 crossbars + 2 discrete memristors

Summary

- Selected publications
  - ICCAD'13, TNNLS'14, ASP-DAC'14, ISCAS'14, IJCNN'14, CogSIMA'14, SSCI'14, SiPS'14, FCCM'15, DAC'15
- Future works
  - HW/SW co-design platform
  - SW: Design a smaller scale representative application for hardware prototyping
  - HW: Improve the scale of NCA design and evaluate its use in larger applications