



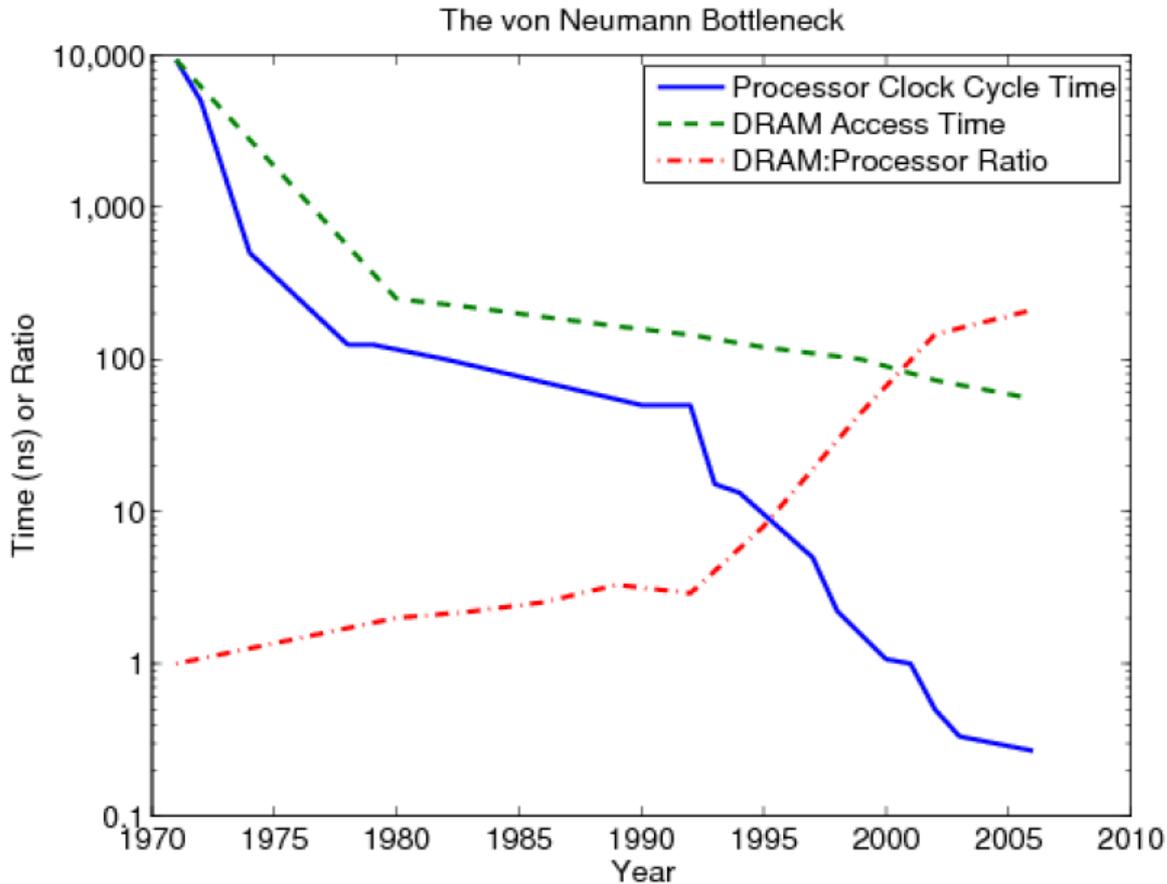
Massively Parallel Cognitive Memory Processor

*“Beyond von Neumann”
Sandia Labs Feb 26th*

Bruce McCormick

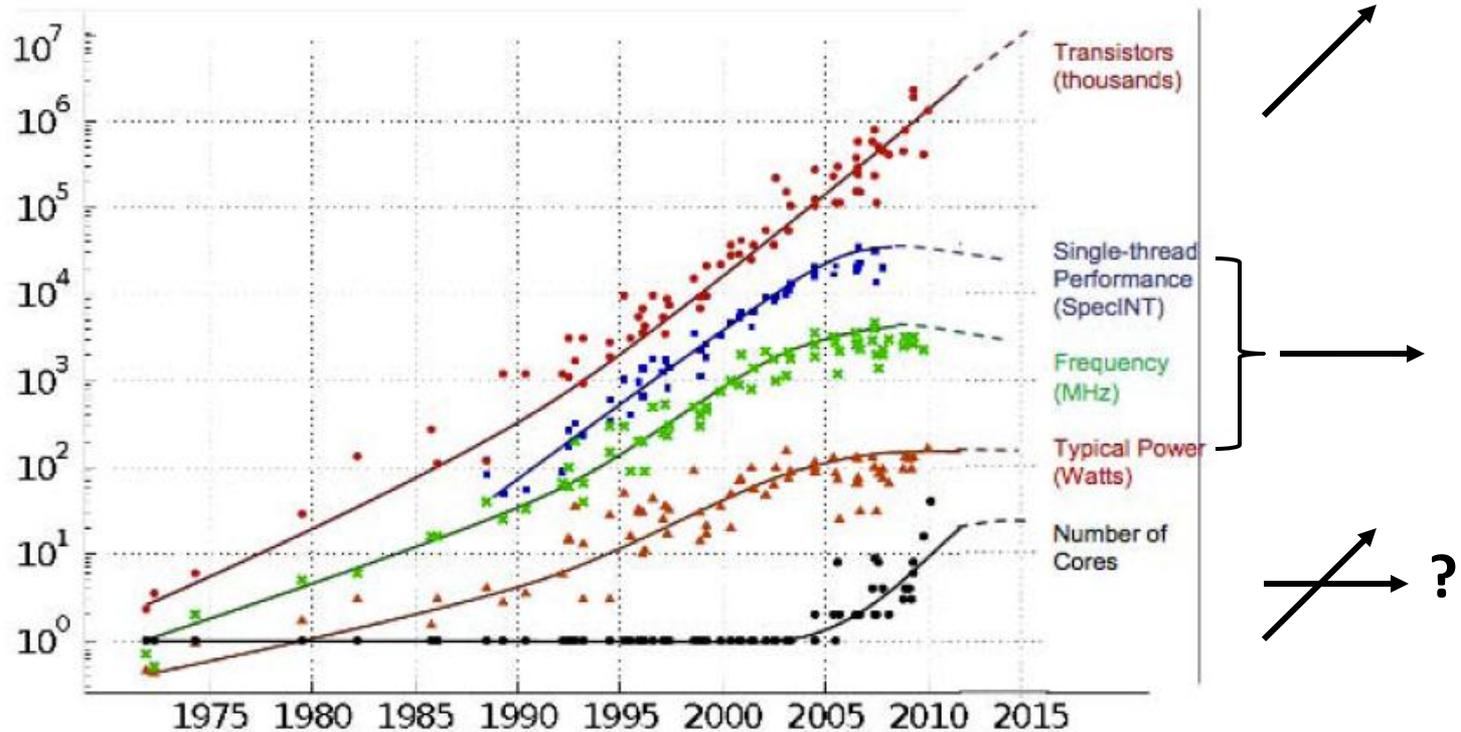
COGNiMEM
Technologies, Inc.

Processor/Memory bottleneck has hit "The Memory Wall"



Meanwhile.....

The End of Historic Scaling



Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

C Moore, *Data Processing in ExaScale-Class Computer Systems*, Salishan, April 2011

- Computing is undergoing a “quiet” change
- Scaling of transistors no longer results in significant performance/ watt improvement
- Companies are going parallel to “compensate”

But...

Historically serial von Neumann architecture has snags going parallel

- Shared and/or distributed memory issues
- Cache coherency and synchronization
- Communication between processes
- Parallel software programming complexity
- Coordination of resources

Amdahl's law pushes the world to try

$1 / ((1-P) + P/N)$ = Performance boost achievable with ideal parallel processing

-- Historically favored going faster serial vs parallel

Von Neumann Computing has hit the Power Wall and the Memory Wall and the Instruction-Level Parallelism Wall !

There is Another Way:

Last century - 2 computing paradigms were created:

One serial- von Neumann Fetch/ Decode-
separate compute & Memory

John von Neumann '46:
“ideally one would desire an indefinitely
Large memory capacity such that any particular
... [memory] word would be immediately
available

The other one - naturally parallel

Machine learning- adaptive pattern recognition

“Embarrassingly” parallel

Eliminates von Neumann bottleneck

Search and sort done in parallel

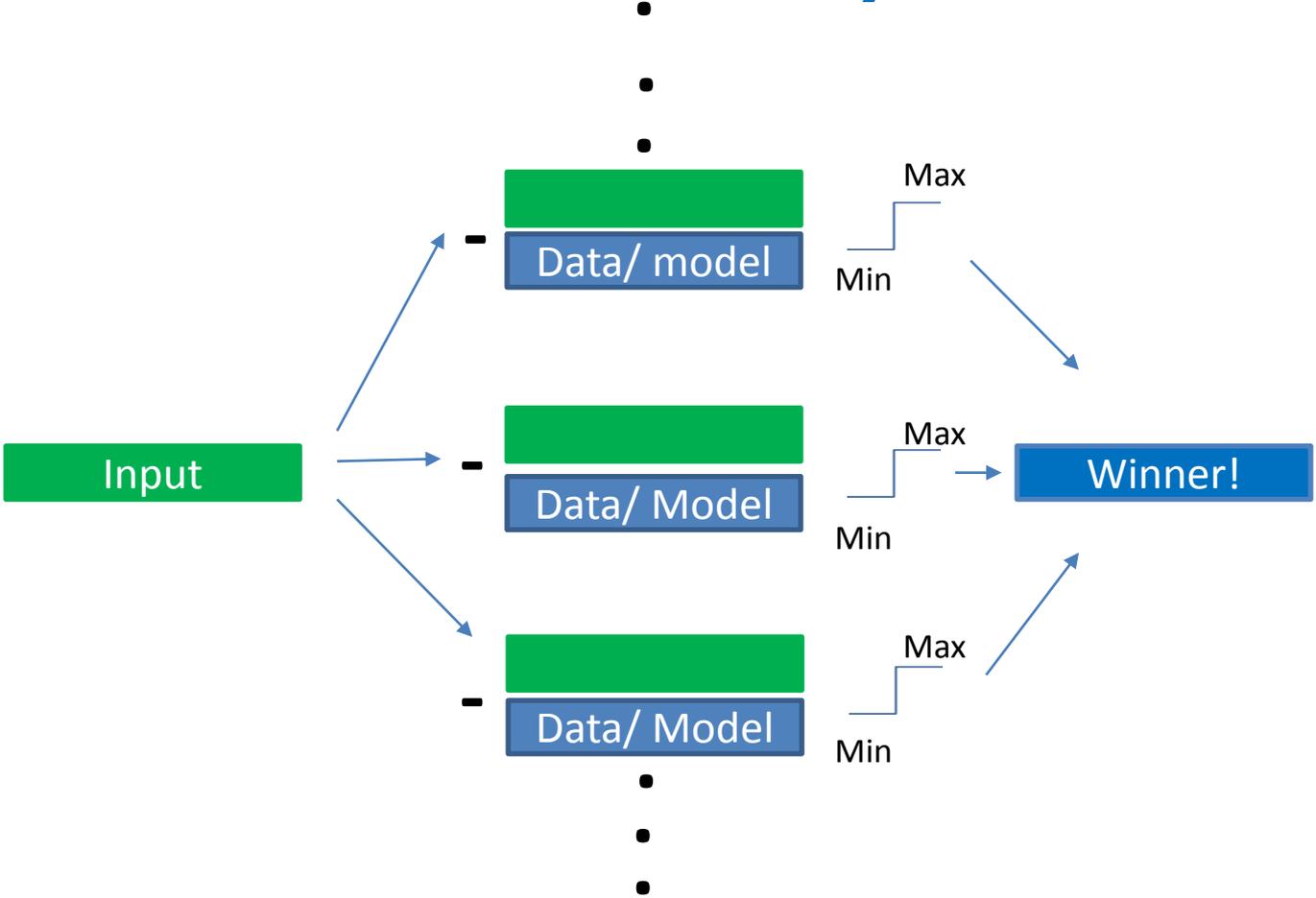
Low power - can run at lower frequencies

Parallel programming/ communication happens “naturally”

This path can help us return to exponential trends in computing

Robert Noyce: “The von Neumann machine may go the way of dinosaurs”-
“random errors will obsolete current logic” “ biological models are
the answer” Jan'85

Cognimem Architecture: 3 layer Network Equivalent



<u>Input</u>	<u>Processing</u>	<u>Classifier</u>	<u>Ordering</u>	<u>Output</u>
≤ 256 Bytes	$\sum_{N=1}^{256} IN - D/M $ or $\text{Max } IN - D/M $	kNN RBF/RCE	Search & Sort	Distance, Category, Exact/ fuzzy Match, No Match, Uncertain

Architecture addresses the parallel problem

Attributes:

- Orders of magnitude lower power
- Eliminates von Neumann memory bottleneck
- Parallel “coding” and communication naturally scale

Applications Examples:

Data clustering through learning/ grouping similar data together

Anomalies are data sets that are “alone”

Biometrics

Pattern recognition- particularly 1/large N – maps directly

Hash function matching (SHA 256 for ex)

Data de-duplication, video fingerprinting

File integrity, validation

Intrusion detection

“Pipeline/ Data Flow Processing”

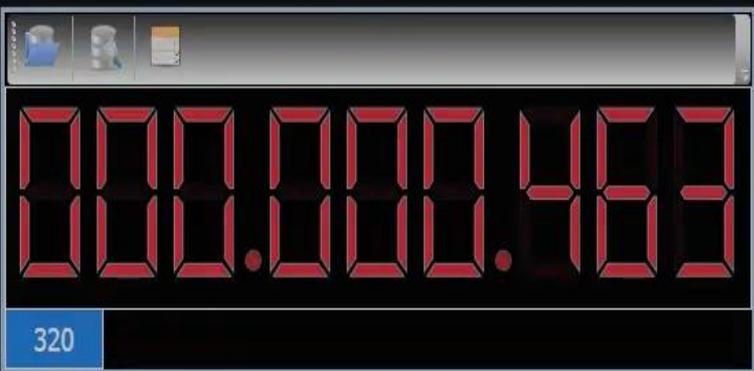
Deep packet processing/ file searching at high transfer rate

DNA searching etc.

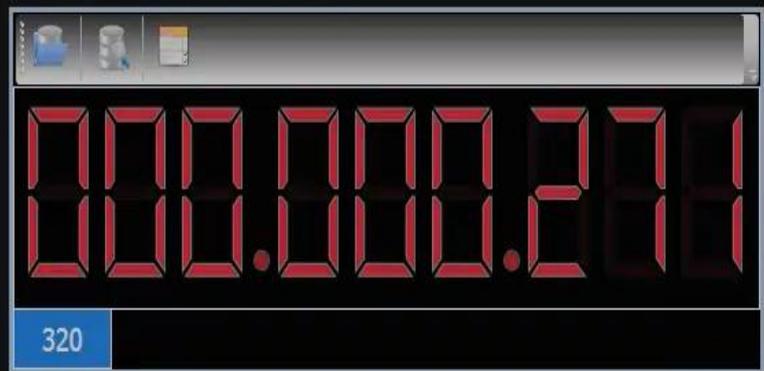
MPEG7

Basically defined to be implemented by NN hardware-ex: compression of frame to frame video by recognizing differences

Platform: Software (sss.mmm.μμμ)



Platform: Hardware (sss.mmm.μμμ)







Tracking Info

Match Prob. (%)

FEAT_DIST:

FEAT_CAT:

NCOUNT:

Calc Time:

Tracking Algorithm

Classic Optimized

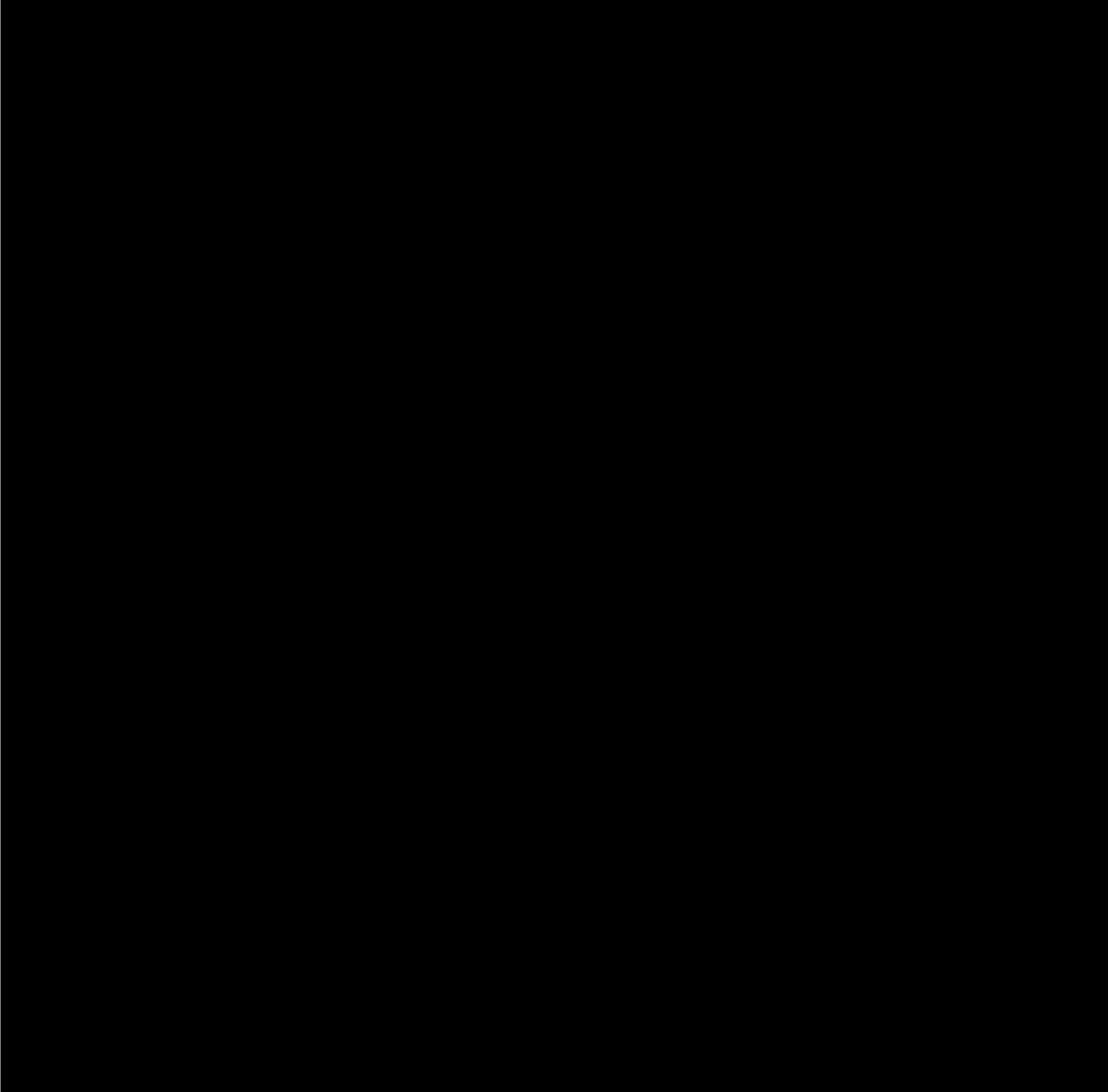
ROI/Tracking Zone

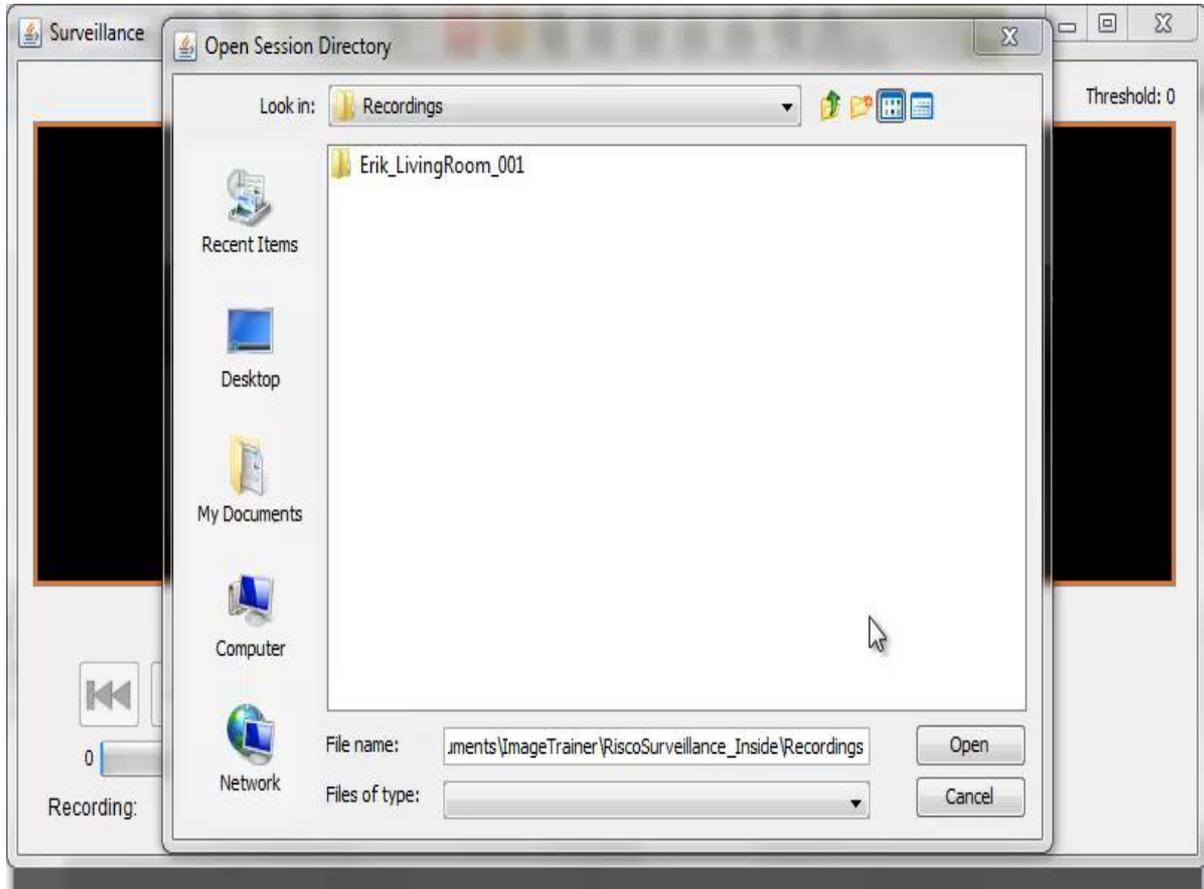
ROI Width ROI Height

TZ Width TZ Height

Tracking Control









10/1/20
November 2012

Recognized Unrecognized

Threshold 1.5000

Interaction

Mouse Control

Interactive Shape

Consecutive Recognitions 1

Security

Record Name recognized as Mouse Up

Mouse Down

Frames

Unrecognized

Record

Save

Max FPS 25



Show details

Show training

愛	液	浴	爭	改	案	以	衣	位	胃	印	失	梅	英	宋
Love	Liquid	Bathing	Fight	Change	Plan	With	Robe	Rank	Stomach	Mark	Lose	Apricot	England	Prosper
置	億	寬	折	貨	課	械	害	街	鏡	各	田	型	固	借
Put	Million	Learn	Break	Money	Impose	Mecha.	Harm	Town	Mirror	Each	Enclose	Type	Solid	Borrow
側	變	完	官	管	閘	視	季	希	紀	器	機	議	給	華
Side	Change	Finish	Government	Tube	Barrier	Observe	Season	Rare	Period	Vessel	Machine	Discussion	Salary	Project
漁	清	共	協	競	極	倉	加	訓	軍	郡	徑	景	藝	欠
Fishing	Pure	Together	Joint	Compete	Pole	Storage	Add	Teach	Army	District	Diameter	Scenery	Art	Lack
健	驗	功	航	娛	康	粉	殺	差	菜	最	材	昨	刷	察
Healthy	Examine	Merit	Cross	State	Healthy	Powder	Kill	Differ	Dish	Most	Material	Before	Print	Guess
參	產	士	氏	史	司	試	兒	治	諱	塩	靜	種	周	祝
Come	Produce	Man	Lineage	Change	Rule	Tri	Child	Rule	Resign	Salt	Quiet	Species	Area	Celebration
順	唱	象	賞	臣	信	巢	末	好	救	成	省	慶	說	節
Order	Advocate	Elephant	Prize	Subject	Believe	Nest	End	Like	Help	Complete	Ministry	Seat	Opinion	Paragraph
選	然	底	卒	帶	隊	戰	達	建	束	單	貯	腸	光	散
Choose	Burn	Bottom	Crisis	Band	Party	Fight	Reach	Build	Tie	Single	Save	Intestines	Sign	Scatter

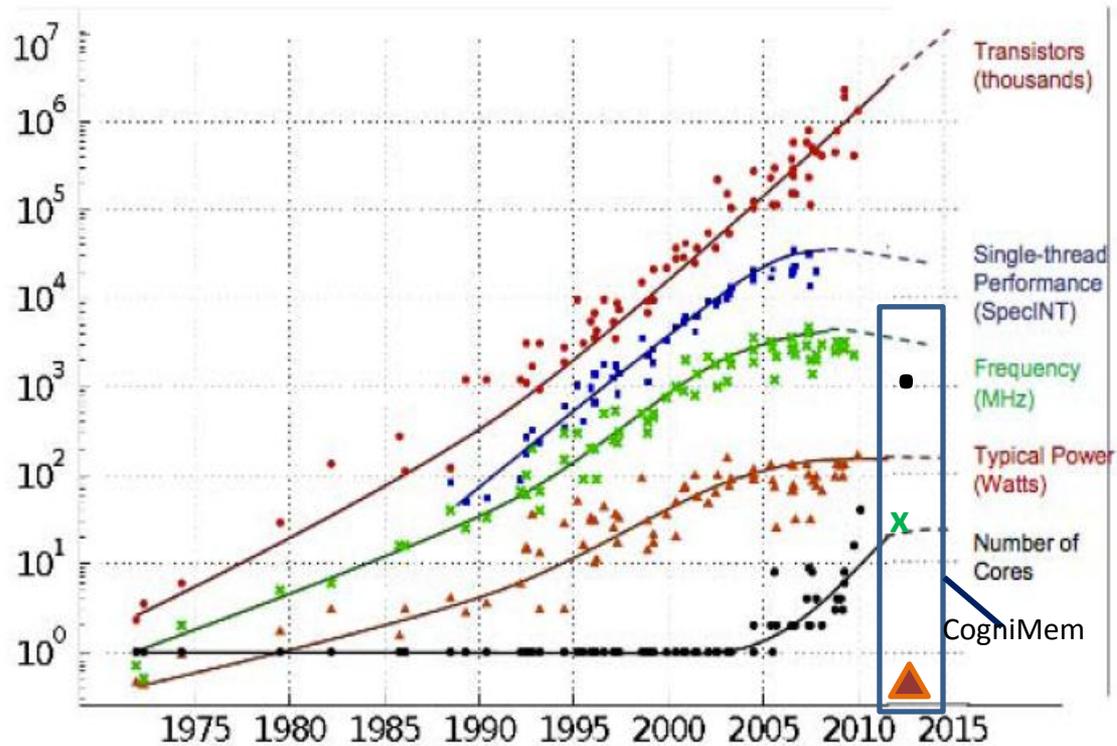


CPU Usage



75%

Cognitive Memory resets the curves



Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

C Moore, *Data Processing in ExaScale-Class Computer Systems*, Salishan, April 2011

Performance scaling extended with the (true) Parallel Architecture

“Beyond” System



1 Million “neurons” in parallel

Target: 1/1million 256 byte match in 10usec

Architecture flexible for:

Applying all neurons onto single task

Partitioning task

< 1 cubic ft

Target : 130 teraops of Pattern Matching Performance ~ 200 watts

Comparison- Honey Bee



A Bee has ~960,000 neurons,
~ 10^9 Synapses

What can a bee do?

“The insects learn to fly the shortest route between flowers discovered in random order, effectively solving the "travelling salesman problem" , said scientists at Royal Holloway, University of London”.

QBI Senior Research Fellow Dr Judith Reinhard said “bees' ‘noses' ... so precise that it could distinguish between hundreds of different aromas and also tell whether a flower carried pollen or nectar, by sniffing its scent from metres away. ”

They can ***FLY, see, navigate, smell, operate in group behavior, etc.***

Future Possibilities

Current architecture can be greatly scaled

Device is on a 130 nm static logic process (vs 22nm advanced, 5 generations)

100K- 1M + neurons/ chip @ 10-30+ x speed

“Beyond” can be 1 chip

1000 gives 1B “processors” in small low power footprint ($\ll 1 \text{ ft}^3$)

Memory technology (memory processing) Non-Volatile options



MRAM, Phase Change Memory, “Memristor” attributes

“As for the human brain-like characteristics- memristor technology could one day lead to computer systems that can remember and associate patterns in a way similar to how people do. This could be used to substantially improve facial recognition technology or to provide more complex biometric recognition systems that could more effectively restrict access to personal information. These same pattern-matching capabilities could enable appliances that learn from experience and computers that can make decisions.”



“IBM researchers unveiled a new generation of experimental computer chips designed to emulate the brain’s abilities for perception, action and cognition. The technology could yield many orders of magnitude less power consumption and space than used in today’s computers.” “The human brain, however, composed of neurons and synapses, does not distinguish between processing (i.e., computation) and memory”.

Parallel processing of a serial stream architectures- “Pipeline/ Data Flow Processing”

Deep packet processing at high transfer rate w/o delay

DNA searching etc.

Summary

As the size of digital information grows, and access increases:

Need more intelligent solutions- for both managing & protecting data
“adaptive, parallel, pattern recognition”

While the current computing paradigm is “running out of gas”
inherently serial & performance scaling fighting physics

It is time:

*For the parallel computing paradigm to augment &
help solve present - future problems!*



Benchmark

Constant 10usec latency of parallel memory access/ match 1 of N
Equivalent to 260 Gops/ watt

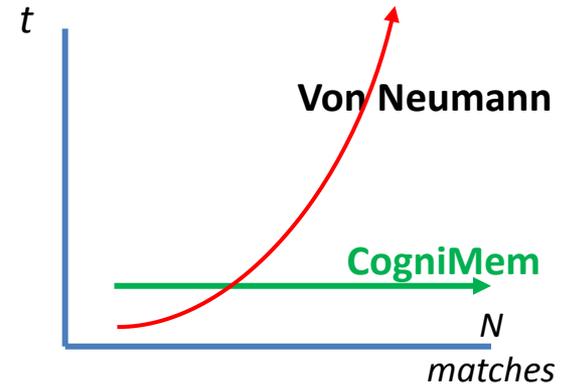
Versus DSP

Significant reduction in green field development time

Performance: 93x tiger shark DSP @ 300mhz*

Performance/watt: up to 330X

*256B against 1024 models of 256B



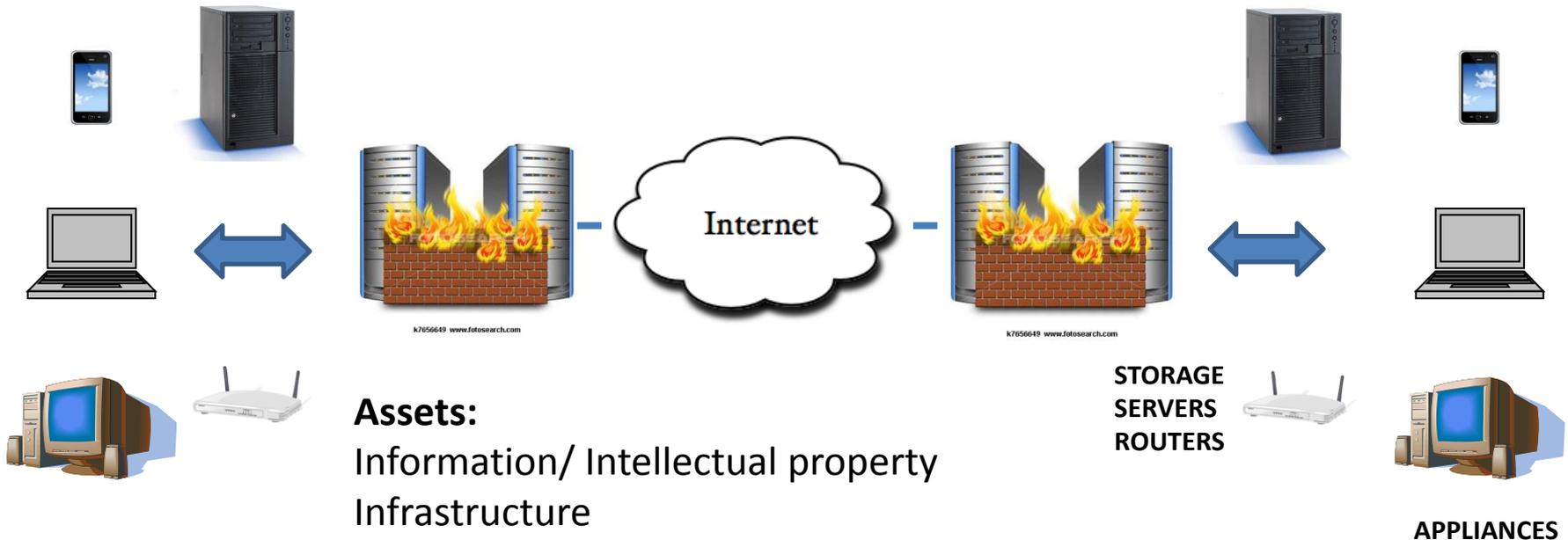
Versus CPU/GPU

Development time reduced vs simulation

Performance	CPU= 1		Cuda (GeForce 8800 GTX)= 1	
	CM1K	Scaled	CM1K (10)	Scaled
KNN	180	3600	3.5	60
	@.5W vs ~65			

CogniMem brings high speed, low power, scalable, general purpose, machine learning and pattern recognition engine

Internet is increasingly a domain asset requiring protection



Threats: - *Level of sophistication of attacks is growing*

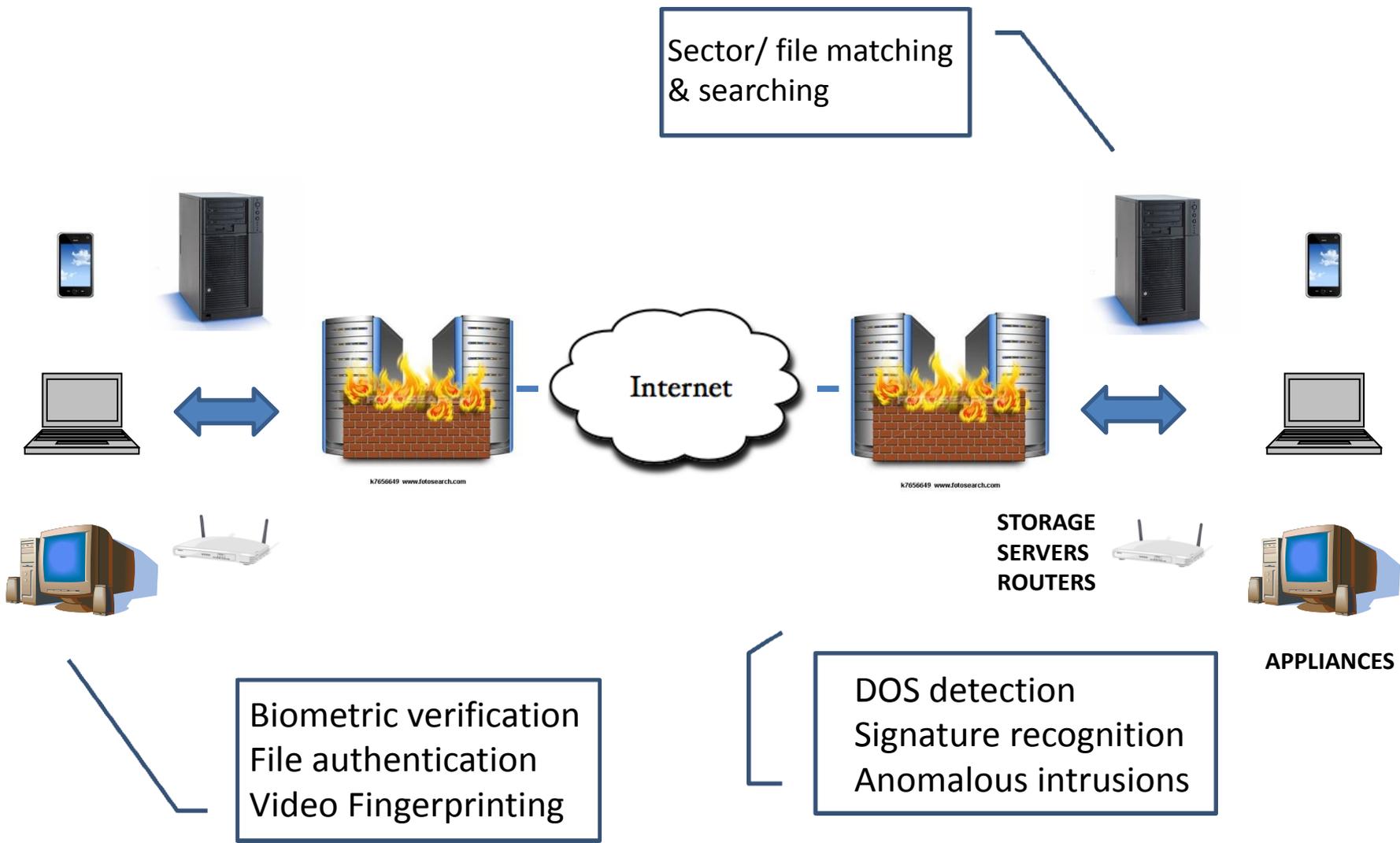
Hacking/ terrorism

Malware, botnets, DOS attacks, theft, Advance Persistent Threats

Corporate Sabotage & Espionage

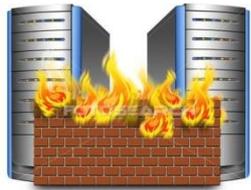
Increasingly powerful tools & solutions are required

Scalable Cognitive Memory can Help



Sector/ file matching
& searching

Internet



STORAGE
SERVERS
ROUTERS

Biometric verification
File authentication
Video Fingerprinting

DOS detection
Signature recognition
Anomalous intrusions

APPLIANCES

Why should Cyber security care?

*Catching anomalous & errant behavior is parallel,
predictive, real time & ultimately adaptive
Pattern Recognition*

Examples:

- Hash function comparison is parallel pattern matching
- Monitoring for malicious “signatures”, DOS is adaptive pattern matching
- Packet inspection needs to be real time, exact and fuzzy
- Biometrics at source and destination needs face, iris, finger, voice recognition etc.
- Clustering of unstructured data for finding similarities/ outliers
- Anomaly detection - comparison of abnormal versus normal patterns

A Hardware Architecture is needed that optimizes Parallel Pattern Recognition