Computing Performance:

*The N3XT 1,000X*

H.-S. Philip Wong

Collaborator: Subhasish Mitra

Department of Electrical Engineering
Stanford University
World Relies on Electronics

Abundant-data
World Relies on Electronics

Internet of Everything
World Relies on Electronics

Genomics
Smart Cities
Military
Health Care
Government
Science
Security
Finance
World Relies on Electronics

Computational demands exceed Processing capability
Many Walls Simultaneously

Power wall

Execution time
- Memory access
- Compute

Memory wall

Complexity wall
- Hardware bugs

Also:
- resilience wall,
- interconnect wall,
- cooling wall, …
THE FUTURE OF COMPUTING PERFORMANCE

Game Over or Next Level?

Samuel H. Fuller and Lynette I. Millett, Editors
Improve Computing Performance

System integration

Device performance
Option 1: Better Transistors

- Few experimental demos
- Transistors ≠ system
Option 2: Design Tricks

- Limited “tricks”
- Complexity → design bugs
Improve Computing Performance

System integration

Multi-cores

Target: 1,000X performance

Business as usual insufficient

Device performance

Power management

Accelerator

Core 1  Core 2  Core 3  Core 4

Core 5  Core 6  Core 7  Core 8
Solution: Nanosystems

Translate new nanotech

New devices

New fabrication

New sensors

imperfections?
large-scale fabrication?
Solution: Nanosystems

Translate new nanotech into new systems

New devices
New fabrication
New sensors

New architectures
Solution: Nanosystems

Translate new nanotech into new systems enable new applications

New devices
New fabrication
New sensors

New architectures
Abundant-Data Applications

Huge memory wall

Application execution time

Compute

96%

Memory access
Computer Chips Today
Limited to 2-Dimensional Circuits
N3XT Nanosystems

Computation immersed in memory
N3XT Nanosystems

Computation immersed in memory

Memory

Ultra-dense vertical connections

Computing logic
N3XT Nanosystems

Computation immersed in memory

Impossible with today’s technologies
Nano-Engineered Computing Systems Technology

[Image of N3XT0101 logo]

[Images of people from Stanford, Berkeley, Carnegie Mellon, and Michigan universities]
Unique N3XT Technology

- End-to-end
  - Isolated improvements inadequate

Existing efforts

- Transistors
- Memories
- New apps
- Multi-core arch. & software
- Chip stacking

N3XT

- 1D / 2D FETs, RRAM, mRAM
- New 3D fabrication
- Nanoscale cooling
- Architecture & software
- Yield, reliability
- Abundant data apps
N3XT Nanosystems

Computation immersed in memory

Computing Logic

Ultra-dense vertical connections
Carbon Nanotube FET (CNFET)

CNT: \( d = 1.2 \text{nm} \)

Energy Delay Product

- \(~ 10X\) benefit
- IBM Power 7 model
CNFET Inverter

INPUT

P+ Doped

N+ Doped

VDD

OUT

GND
Big Promise, Major Obstacles

- Process advances alone inadequate

Solution: Imperfection-immune design

[Zhang IEEE TCAD 12]
CNT Growth *circa* 2005

- Highly mis-positioned
First Wafer-Scale Aligned CNT Growth

Quartz wafer with catalyst

Aligned CNT growth

Quartz wafer with CNTs
99.5% aligned CNTs

Stanford Nanofabrication Facility

[Patil VLSI Tech. 08, IEEE TNANO 09]
Wafer-Scale CNT Transfer

High-temperature CNT growth

Low-temperature circuit fabrication

900 °C

CNT transfer

120 °C

Before transfer

After transfer

Quartz

SiO₂/Si

2 µm

2 µm

[Patil VLSI Tech. 08, IEEE TNANO 09]
Mis-Positioned CNT-Immune NAND

1. Grow CNTs

2. Extended gate, contacts

3. Etch gate & CNTs

4. Dope P & N regions

- Arbitrary logic functions
  - Graph algorithms

Etched region essential
VLSI Metallic CNT Removal

- Chip-scale electrical breakdown
- Universally effective

99.99% m-CNT removal, 4% s-CNT removal

[Patil IEDM 09, IEEE TNANO 10, Shulaker ACS Nano 14]
New VMR

- Arbitrary technology nodes: 10nm & beyond

Relaxed node $\rightarrow$ m-CNTs **Erased** $\rightarrow$ Scaled circuits

Record selectivity

99.99% m-CNTs erased, 1% s-CNTs erased
Most Importantly

- VLSI processing
  - No per-unit customization

- VLSI design
  - Immune CNT library
First Sub-system: ISSCC Demo
First Sub-system: ISSCC Demo

ISSCC Jack Raper Outstanding Technology Directions Paper

Sacha: CNT Controlled Hand-shaking Robot

Wafer with CNFET circuits

Robot

CNT Computer

[Shulaker Nature 13]
CNT Computer

- Turing-complete processor: entirely CNFETs

[Diagram of CNT Computer with Instruction Fetch, Data Fetch, ALU, and Write-back sections]

[Shulaker Nature 13]
System Demos

Video:
https://www.youtube.com/watch?v=7lmK4iNrlGo&feature=youtu.be
Reproducible Results

80 ALUs

200 D-Latches

~ 1,600 CNFETs

~ 1,800 CNFETs

Waveforms overlaid
High-Performance CNFETs

Doping

Current Drive

Dielectric interactions

Contact Resistance

Scaling
High-Performance CNFETs

- > 100 CNTs/µm
  - Major challenge
- New result
  - > 100 CNTs/µm
  - Record $I_{ON}$ density
  - Controlled variations

[Shulaker IEDM 14]
High Performance Obstacles

- Doping
- Current Drive
- Dielectric interactions
- Contact Resistance
- Scaling
Complementary CNFET Logic

- VDD from 1.0V → 0.8V → 0.6V → 0.4V → 0.2V

• n-CNFET
• p-CNFET

Drain Current (A) vs Gate Voltage (V)

L = 2 μm
W = 2 μm
SS = 95 mV/dec

L = 90 nm
W = 250 nm
SS = 97 mV/dec
High Performance Obstacles

Doping
Current Drive
Dielectric interactions
Contact Resistance
Scaling
Recent Progress

Top-contact

Edge-contact

[Cao, Science 15 (IBM)]
N3XT Nanosystems

Computation immersed in memory

Memory

Computing Logic

Ultra-dense vertical connections
Many Nano-scale Innovations

Memory & logic devices

3D Resistive RAM (RRAM)

MoS$_2$

2D FETs: large-area monolayer MoS$_2$

Embedded cooling

Phase change: hotspots suppressed

Vertical metal nanowire arrays

Metal pillar electrode
Memory cell
SL
Vertical MOSFET
BL
Metal plane electrode (WL)

Evaporative Wicking
30 µm thick

Power Density (W/cm$^2$)

Temperature (°C)

L=109.87 um

L=5 um
New Memories

Random access, non-volatile, no erase before write

STT-MRAM
Spin torque transfer magnetic random access memory

PCM
Phase change memory

RRAM
Resistive switching random access memory

CBRAM
Conductive bridge random access memory
Scalable Embedded Memory

Bi-layer TiOx (2.5nm) / HfOx (1.5nm)

Scalable: 12 nm

Scalable: 10 nm


B. Govoreanu et al., *IEDM* 2011 (IMEC)
High Density 3D Memory

Stanford:
IEDM ’12, ’13,
VLSI ’13, ’14,
DATE ’15, Nature Comm ‘15
High Density 3D Memory

- < 1 μA
- 1 – 2 V
- 5 ns
- > 1G cycles
- F = 5 nm
- 128 layers
- 64 Tb per chip

Stanford:
IEDM ’12, ’13,
VLSI ’13, ’14,
DATE ’15, Nature Comm ‘15
N3XT Nanosystems

Computation immersed in memory

Ultra-dense vertical connections

Memory

Computing Logic
The “High-rise” chip

Circuit demos

Routing Element

WL[0]

BL_1

Routing Element

WL[1]

BL_2

Routing Element

WL[2]

Routing Element

WL[3]

Logic

RAM

RAM

Logic

Si-FET

RRAM

CNFET

200 µm

[Shulaker IEDM 14]
Interwoven Compute + Memory + Sensing

Terabytes / second

Millions of sensors

Ultra-dense vertical connections

To be published. Please keep in confidence

Abundant sensor data:
Extensive, accurate classification

[M. Shulaker, Stanford. Unpublished]
Complement with Software Solutions

Co-optimized s/w + h/w

Runtime optimization

Yield, reliability

Learning: key architectural concept

Cross-Layer Resilience

DSL compiler

DSL = Domain-Specific Language
Quantifying N3XT System Benefits

- Heterogeneous nanotechnologies
- Architecture design space
- Physical design
- Integrated thermal analysis
- Yield, reliability
Sweet Spot: Abundant-Data Apps.

IBM graph analytics
DeepDive
BigDataBench
Data-intensive computing

851X benefits

Energy: 37X
Exec. Time: 23X

PageRank app.
Sweet Spot: Abundant-Data Apps.

IBM graph analytics

DeepDive

BigDataBench

Data-intensive computing

851X benefits

Energy: 37X

Exec. Time: 23X

Processors active

Processor idle

Memory access

PageRank app.
Massive Benefits Require

- Not a logic device
- Not a memory device
- Not 3D integration
- Not thermal management
- Not new architectures
- Not yield and reliability management
Massive Benefits Require

- a logic device
- a memory device
- 3D integration
- thermal management
- new architectures
- yield and reliability management
Sponsors

Stanford  SystemX Alliance

Non-Volatile Memory Technology Research Initiative
Conclusion

- **Nanosystems today**

- **Game ON, to the era**

- **N3XT 1,000X**
  - Compute + memory densely interwoven
N3XT Nanosystems

Computation immersed in memory

Memory

Ultra-dense vertical connections

Computing logic