

Superconducting Architecture

CRNCH SUMMIT 2020

Brian Konigsburg
January 31, 2020

This research is based upon work supported by the ODNI, IARPA, via ARO contract number W911NF-14-C-011x. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government.

Opportunity for Advanced Computing

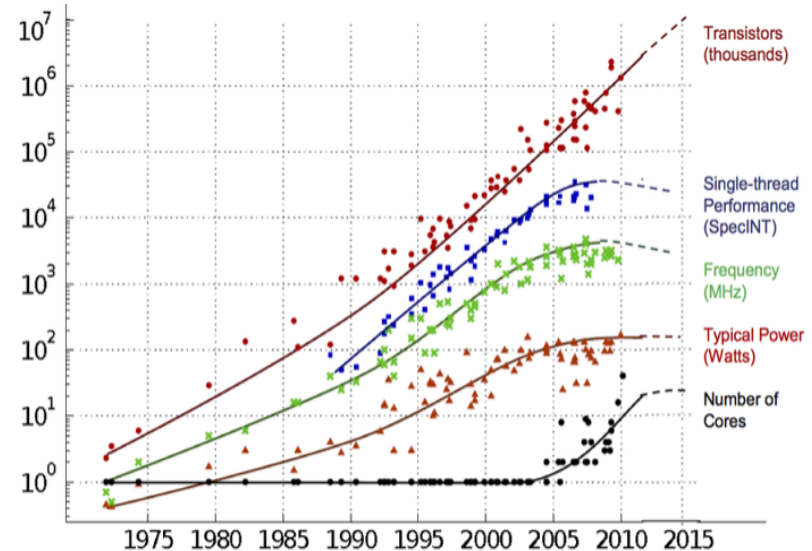
Technological Improvements are Needed Today!

Silicon Computing Challenges

- Diminishing returns in node changes
- Power, heat, clock and dark silicon limitations
- Economic hurdles to state of the art

RQL for Beyond Silicon Computing

- Reciprocal Quantum Logic
- 50-300X improvement in FLOPS/watt
- Suitable for supercomputing & data centers
- Provides a path beyond exascale



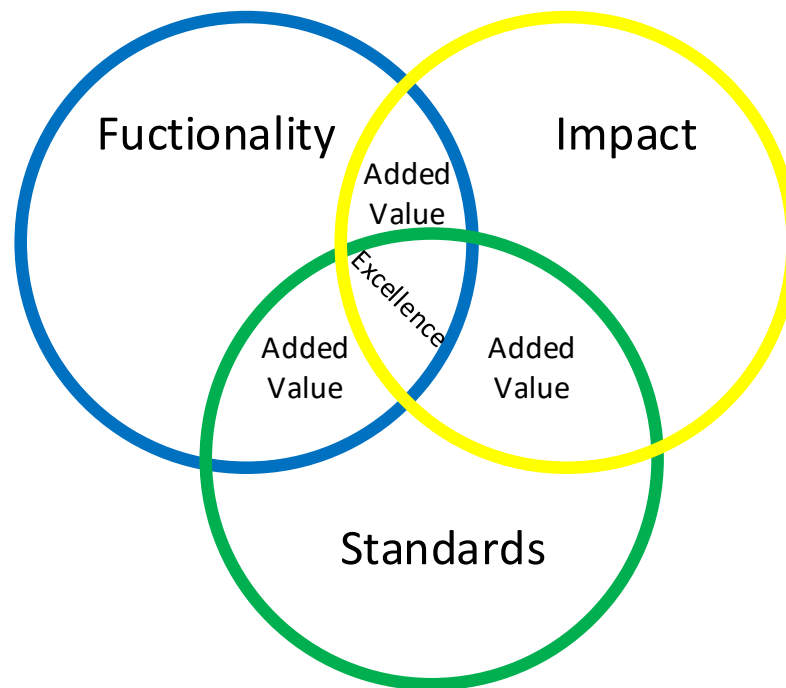
Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

RQL Addresses CMOS Power Issue

Superconducting Architecture Goals

Questions to answer

- Re-Map or Re-Think
- Technology's influence on Architecture
- RQL vs CMOS metrics
 - Area, performance, power
- Prototype RQL architecture
- Analysis of Architecture

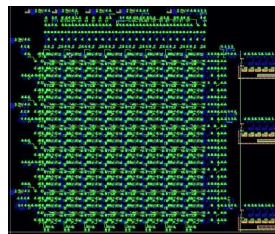
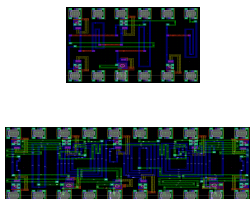


What Does an Architecture tuned to RQL Look Like?

RQL Digital Progression

3 years of Continuous Improvement in Design, Fab and Test

NGMS IR&D

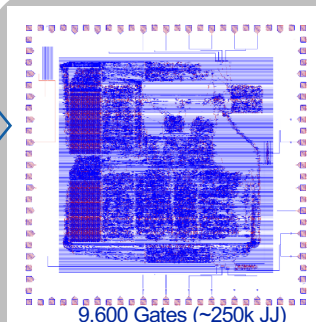
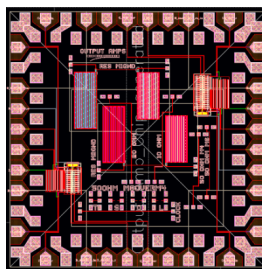
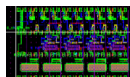


~1,000 Gates (~27k JJ)

Gates, Amplifiers,
and Shift Registers

Components

IARPA C3

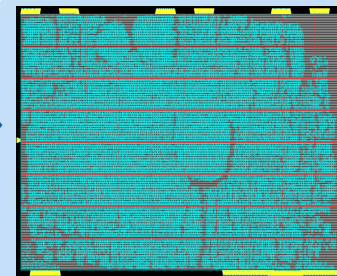


9,600 Gates (~250k JJ)

Limited 32-bit

8bit CPU

5,600 Gates (~50k JJ)

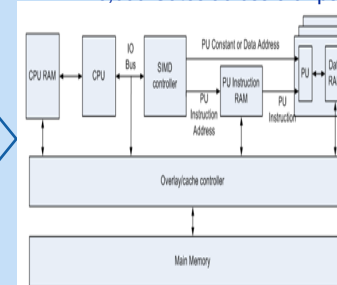


18,300 Gates (~495k JJ)

Full 32-bit

16bit SIMD

75,000 Gates across 5 chips



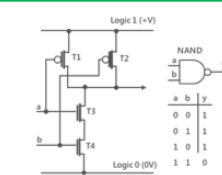
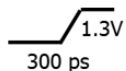
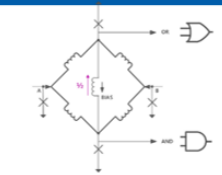
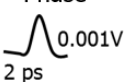
Progress being made on circuit density and functionality



RQL for Architects

Basic RQL Properties

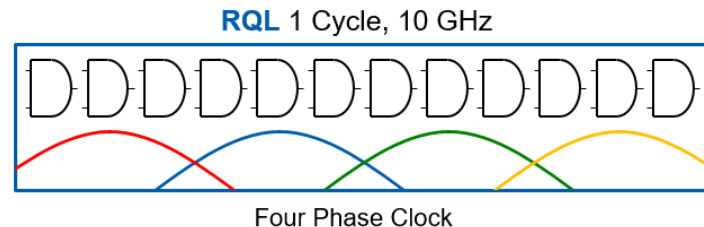
Lossless, fast wires
 Non-inverting basic gate
 Low switch power
 Low gate gain (fanout)
 Resonator clock
 Power from AC clock
 Requires low temp

CMOS	Data Encoding	Active Element	Performance	Gate Density	Wiring
	Voltage 	n- and p-type Silicon transistors 10 ⁻¹⁵ Joules Switching Energy	Driven by <u>Litho</u> Node, Supported by Material	Driven by <u>Litho</u> Node	Resistance & Capacitance Driven Energy Intensive
RQL	Data Encoding	Active Element	Performance	Gate Density	Wiring
	Current Pulse or Junction Phase 	Superconductor -Insulator Tunnel Junction 10 ⁻¹⁹ Joules Switching Energy	Driven by Materials NOT <u>Litho</u> Node	Driven by metal-insulator Layer Count and <u>Litho</u> Node	Inductance Driven Lossless Transmission

Significant Fundamental Differences from CMOS

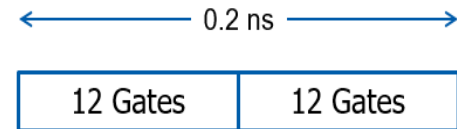
RQL Layout and Timing

- Gates compatible with tools
- Gate performance relative to resonator power
 - More phases, better perf
- Latch based timing
- No physical latch boundary
- Supports 2-4x frequency



RQL

- 12 Gates per clock @ 10 GHz



**Logic
Depth**

24

CMOS

- 24 Gates per clock @ 5 GHz



24

RQL Design Methodologies

Return to Zero (Wave Pipeline)

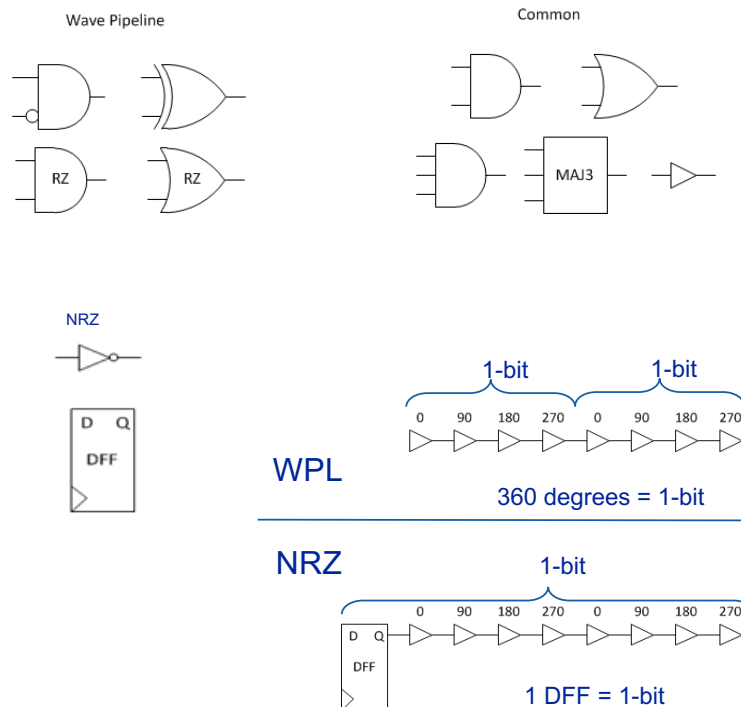
- Very high throughput
- Data always moving
- Runs at AC clock/power rate

Non-Return to Zero (NRZ)

- CMOS-like design style
- Traditional controllable clock
- Lower Logic Power (sometimes)

Interfacing is relatively simple

- Hybrid design best of both approaches



Wave pipeline Inverter

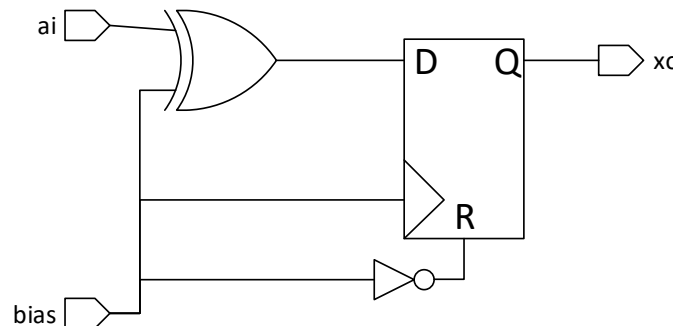
WPL inverter based on an XOR

Equivalent circuit for what it actually does

Means an AC phase boundary is needed
at every one of these

More phases means less penalty waiting
for boundary

NRZ has true inverter



Interconnect: Fast and Virtually Free

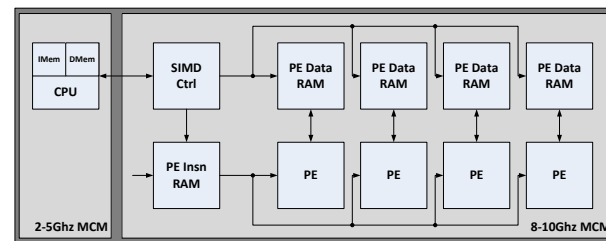
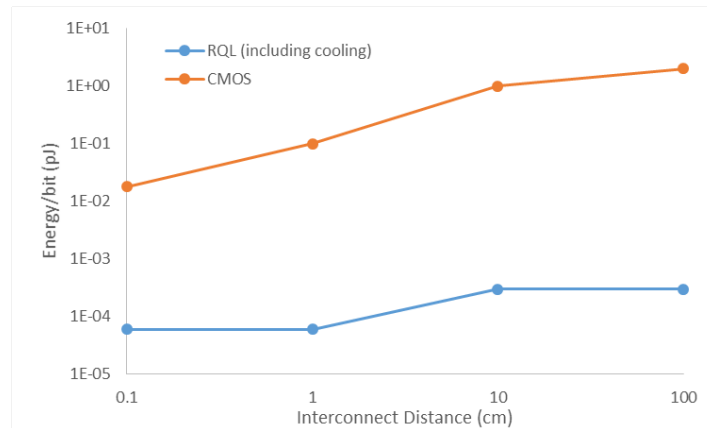
Large superconducting computers competitive in density per **volume** rather than area

RQL feature size drive multi-chip implementations so MCM support is critical

Need PTLs to efficiently communicate between chips on MCM

PTLs between blocks avoid timing penalties that reduce performance

Efficient, low latency long distance wires - major advantage over CMOS



Superconducting Memories

NDRO

- Large Cell (device limited)
- Timing per bit doesn't support high performance for large arrays

77K DRAM

- DRAM technology for high density
- 77K for lower power
- Good bulk memory solution

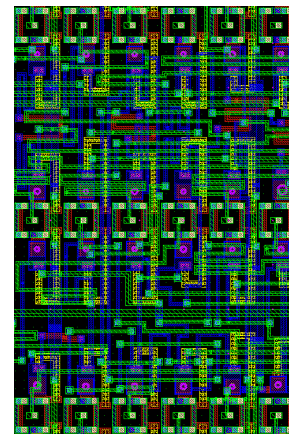
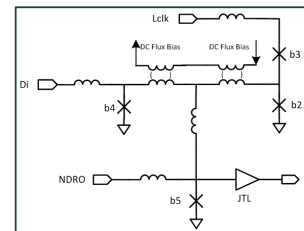
Other technologies being explored

- SRAM like timing and density under development (L1/L2 cache)
- Magnetic RAM for high density (L3 cache)

Faster wires mean smaller caches possible

Low power wires mean more, wider busses possible

NDRO RAM



RQL Compute Density

100x compute density regardless of architecture at large scale

- Applies to CPUs/GPUs/FPGAs or special purpose processors

CMOS 3U rack mount vs RQL

- 3U spacing for heat dissipation
- RQL spacing: 7-8 mm
- >15x volume advantage

3D stacking

- Favorable RQL thermal properties

1000x CMOS wire power vs RQL

- 1.3V CMOS vs 0.001V RQL
- Power function of V^2

	Initial Node		Future Node	
	CMOS	RQL	CMOS	RQL
Processor	244.0	1	165.0	1
Chip/MCM	2.6	1	1.8	1
Board	1.3	1	1.0	2
Rack	1.0	96	1.0	142

Architectural Opportunities

Point-to-point vs Shared busses

- RQL wire power, delay advantage

Frequency vs Performance

- Data flow, Pipelines 10+ GHz

Single Chip vs Multi-chip

- Chip-to-Chip penalties lower in RQL

Memory Hierarchy

- Size vs Speed trade-off

Fanout

- MUX Selects, High fanout control, reset

Inversion

- Dual rail, Synthesis reduction

- Routing
 - JTLs more dense vs PTLs faster
- Wave-pipelined vs NRZ
 - Balance performance and area/power
- Revisit CMOS Axioms
 - Pipeline FIFOs rather than stalls
 - Crossbar switches rather than shared busses
 - Data flow rather than control flow

Conclusions

Attributes to take advantage of:

- At speed, wave-pipelined nature of RQL means we get clock scaling back
- Nearly free resistive interconnects means high cross sectional bandwidth networks are now a large routing problem
- Scaling to large (160x240mm) “systems on a chip” vs multi-chip for dense compute capability at much lower power
- Non-von Neumann architectures – spiking neural networks – seem like a natural fit for single flux quanta pulses

Heterogeneous computing is also possible

- Similar to what we see now with GPUs, FPGAs, Xeon Phi, etc.

A new era of computer architecture required for RQL

- From small A – micro-architecture to big A – system level architecture

NORTHROP
GRUMMAN

The logo graphic consists of a thick horizontal line extending from the end of the word "NORTHROP" to the right, and a thick vertical line extending downwards from the end of the word "GRUMMAN". These two lines meet at a right angle, forming an L-shape that frames the top-right corner of the text.