National Strategic Computing Initiative Seminar Program
June 7, 2016 – Feb 7, 2017

NSF Stampede
U of Texas/Austin

DoE Titan/ORNL

NSF Blue Waters/NCSA
NSCI

On July 29 Obama signed and EO creating the National Strategic Computing Initiative. In order to maximize the benefits of HPC for economic competitiveness and scientific discovery, the United States Government must create a coordinated Federal strategy in HPC research, development, and deployment. Investment in HPC has contributed substantially to national economic prosperity and rapidly accelerated scientific discovery. Creating and deploying technology at the leading edge is vital to advancing my Administration's priorities and spurring innovation. Accordingly, this order establishes the National Strategic Computing Initiative (NSCI).
NSCI

• Started as whole of govt effort and now is a whole of nation effort.
• DoE, DoD and NSF are lead agencies
• NIST and IARPA are foundational research agencies
• NASA, FBI, NIH, DHS and NOAA are deployment agencies
  • Each chartered to support the NSCI in ways consistent with the mission and traditional strengths of that agency but to work together to ensure that the US moves forward and to avoid duplication
• Importance of NSCI
  • Scientific discovery and economic competitiveness required action
• Many complex scientific problems rely on both simulation and/or the manipulation, curation and analysis of very large data sets.
• Separations of these two aspects are being blurred and new approaches are required if we are to move toward from petascale to exascale.
• While NIST will play a central role as a foundational agency, there was a clear need both internally and externally to bring current expertise to educate and also to help formulate the path to achieving the NSCI goals.

• **NIST decided to start an NSCI seminar program.**

  **NSCI Internal Web Site**

  **Organizing Group:** Vladimir Aksyuk, Albert Davydov, Bob Hickernell, Curt Richter and Barry I. Schneider  
  **Support:** Hoyt Cox (AV), Zulma Lainez (G‘berg) and Kimberley Keinath (Boulder)

• The initial thinking was that it would be only an internal seminar program but as a consequence of much broader interest, it was decided to allow outsiders to attend as visitors, subject to the usual NIST rules and to webcast the talks to selected sites. I am happy to say that this has been accomplished.
Kickoff Talk:

The IARPA Cryogenic Computing Technology program

Marc Manheimer
Program Director IARPA Superconducting Computer Program.
Superconducting Computing

Marc Manheimer
IARPA Program Manager
7 June 2016
NIST NSCI Seminar
Superconducting Computing

- Cryogenic Computing Complexity Program Motivation
- Program Snapshot
- Technical Background
- Why this is hard
- Miscellaneous
C3 Program Motivation
Our appetite for computing is unlimited!

• Upgrading a facility to more powerful computers is constrained by
  – **Power** supply capability of electric company
  – **Space** limitations
  – **Cooling** infrastructure

• Constraints on developing computers with additional processing power
  – Some estimates to reach exascale are in the hundreds of megawatts.
  – An exascale computer at 20 megawatts based on semiconducting technology will require heroic measures.
  – We will require a different technology to get beyond exascale.

* A computer based on *superconducting logic* and *cryogenic memory* can help solve these issues
Problem: Increasing Power Requirements For Conventional Supercomputers
Superconducting computing looks promising

Key Factors

• Approach based on:
  – *Near-zero energy* superconducting interconnect
  – *New* SFQ logic with no static power dissipation
  – *New energy efficient* cryogenic memory ideas
  – *Electrical or optical* ingress/egress
  – *Commercial* cryogenic refrigerators

**IARPA C3 program basis**
# Conceptual System Comparison (~20 PFLOP/s)

## Titan at ORNL vs. Superconducting Supercomputer

<table>
<thead>
<tr>
<th></th>
<th>Titan at ORNL</th>
<th>Superconducting Supercomputer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Performance</strong></td>
<td>17.6 PFLOP/s (#2 in world*)</td>
<td>20 PFLOP/s</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>710 TB (0.04 B/FLOPS)</td>
<td>5 PB (0.25 B/FLOPS)</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>8,200 kW avg. (not included: cooling, storage memory)</td>
<td>80 kW total power (includes cooling)</td>
</tr>
<tr>
<td><strong>Space</strong></td>
<td>4,350 ft² (404 m², not including cooling)</td>
<td>~200 ft² (19 m², includes cooling)</td>
</tr>
<tr>
<td><strong>Cooling</strong></td>
<td>additional power, space and infrastructure required</td>
<td>All cooling shown</td>
</tr>
</tbody>
</table>

* #1 in TOP500, 2012-11 (17.6 PFLOP/s)
C3 Program Snapshot
C3 Notional System, Metrics, and Goals

### Cryogenic Refrigerator

- Host
- Input/Output
- CPU
- PA Controller
- Parallel Accelerator
- Cache
- Main Memory
- Room Temperature
- ~4 kelvins (-270 °C)

#### Metrics and Goals

<table>
<thead>
<tr>
<th>Metric</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clock rate for superconducting logic</td>
<td>10 GHz</td>
</tr>
<tr>
<td>Throughput (bit-op/s)</td>
<td>$10^{13}$</td>
</tr>
<tr>
<td>Efficiency @ 4 K (bit-op/J)</td>
<td>$10^{15}$</td>
</tr>
<tr>
<td>CPU count</td>
<td>1</td>
</tr>
<tr>
<td>Word size (bit)</td>
<td>64</td>
</tr>
<tr>
<td>Parallel Accelerator count</td>
<td>2</td>
</tr>
<tr>
<td>Main Memory (B)</td>
<td>$2^{28}$</td>
</tr>
<tr>
<td>Input/Output (bit/s)</td>
<td>$10^9$</td>
</tr>
</tbody>
</table>
### C3 Phase 1 Metrics and Goals

<table>
<thead>
<tr>
<th>Metric</th>
<th>BP</th>
<th>OP1</th>
<th>OP2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cryogenic Memory</strong></td>
<td>Memory cell</td>
<td>Array</td>
<td>Chip</td>
</tr>
<tr>
<td><strong>Functional capacity (bit)</strong></td>
<td>1</td>
<td>$2^6$ ; $2^6$</td>
<td>$2^{10}$ ; $2^{10}$</td>
</tr>
<tr>
<td><strong>Density (bit/cm$^2$)</strong></td>
<td>$10^6$ ; $10^5$</td>
<td>$5 \times 10^6$ ; $5 \times 10^5$</td>
<td>$10^7$ ; $10^6$</td>
</tr>
<tr>
<td><strong>Data rate, burst mode (Gbit/s)</strong></td>
<td>1</td>
<td>5 ; 30</td>
<td>5 ; 30</td>
</tr>
<tr>
<td><strong>Access time, ave. (ps)</strong></td>
<td>10,000 ; 1,000</td>
<td>5,000 ; 400</td>
<td>5,000 ; 400</td>
</tr>
<tr>
<td><strong>Access energy, ave. (J/bit)</strong></td>
<td>$5 \times 10^{-16}$ ; $5 \times 10^{-17}$</td>
<td>$5 \times 10^{-16}$ ; $5 \times 10^{-17}$</td>
<td>$10^{-16}$ ; $10^{-17}$</td>
</tr>
<tr>
<td><strong>Logic, Comm. &amp; Systems</strong></td>
<td>Subcircuits</td>
<td>Circuits</td>
<td>Processors</td>
</tr>
<tr>
<td><strong>Benchmark circuits &amp; applications</strong></td>
<td>Circuits 1</td>
<td>Circuits 2</td>
<td>Circuits 3</td>
</tr>
<tr>
<td><strong>Complexity (JJ)</strong></td>
<td>$10^4$</td>
<td>$5 \times 10^4$</td>
<td>$10^5$</td>
</tr>
<tr>
<td><strong>Density (JJ/cm$^2$)</strong></td>
<td>$10^5$</td>
<td>$5 \times 10^5$</td>
<td>$10^6$</td>
</tr>
<tr>
<td><strong>Throughput (bit-op/s)</strong></td>
<td>$10^9$</td>
<td>$5 \times 10^{10}$</td>
<td>$10^{11}$</td>
</tr>
<tr>
<td><strong>Efficiency @ 4 K (bit-op/J)</strong></td>
<td>$10^{16}$</td>
<td>$5 \times 10^{16}$</td>
<td>$10^{17}$</td>
</tr>
</tbody>
</table>

* Memory metrics: The first number refers to Main Memory and the second to Cache Memory.
C3 Organization

Government

Performers teams

Phase 1

LCS Thrust
Logics, Communications & Systems

CM Thrust
Cryogenic Memory

Phase 2

Logic, Communications & Systems

Chips

NIST Test and Evaluation

SNL Failure Analysis
C3 Augmentations

- IARPA SuperTools program – BAA is in preparation

- Official Request For Information on data ingress and egress
  - Published on FedBizOpps; responses received in March
  - Likely new program
## Performer Teams

### Cryogenic Memory

<table>
<thead>
<tr>
<th>Performer</th>
<th>Institution 1</th>
<th>Institution 2</th>
<th>Institution 3</th>
<th>Institution 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northrop Grumman Corporation</td>
<td>Arizona State University</td>
<td>Michigan State University</td>
<td>Raytheon BBN Technologies</td>
<td>BBN Technologies</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cornell University</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Massachusetts Institute of Technology</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>New York University</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>University of Rochester</td>
<td></td>
</tr>
</tbody>
</table>

### Logic, Communications, & Systems

<table>
<thead>
<tr>
<th>Performer</th>
<th>Institution 1</th>
<th>Institution 2</th>
<th>Institution 3</th>
<th>Institution 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM Watson Research Center</td>
<td>Hypres Inc.</td>
<td>Rensselaer Polytechnic Institute</td>
<td>Stony Brook University</td>
<td></td>
</tr>
<tr>
<td>Northrop Grumman Corporation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Government Teams

Foundry

MIT Lincoln Lab

Test & Evaluation

NIST Boulder

Sandia National Labs

Contracting In Process
Technical Background
Superconductors

- Current flows without resistive losses below a critical temperature

- Electrical current in a loop stores flux
  - Can be basis for memory, $I \cdot L = \Phi$
  - Flux is quantized; minimum value is $\Phi_0 = h/2e \approx 2.07 \text{ fWb (mV} \cdot \text{ps or mA} \cdot \text{pH)}$
  - To hold one flux quantum, a 20 micrometer diameter niobium loop would carry about 30 microamperes

- Supercurrent can flow across a thin insulating barrier between two superconductors: Josephson tunneling
  - This works up to a critical current, $I_c$
  - For $I>I_c$, an ac voltage develops across the barrier
  - For a short input pulse $\int Vdt = \Phi_0$
  - Typical pulses are a picosecond long and millivolts tall, with energy $\sim 10^{-19}$ joules
Simple Circuits: NAND and RS Flip-flop
COST Memory (Raytheon-BBN, NYU)

- Stands for Cryogenic Orthogonal Spin Transfer
- Requires multiple magnetic layers (MTJ structure)
- Electron current is spin polarized in passing through permanent magnet polarizing layer
- Polarized electrons exert torque on switchable magnetic layer
- Magnetization of switchable layer is reversed
- This changes resistance of stack.
Cryogenic Spin Hall Effect Memory (Raytheon-BBN, Cornell U)

- Spin orbit coupling of the electron current in the spin-Hall channel gives rise to an orthogonal spin current that goes into the magnetic tunnel junction (MTJ) structure.
- The spin current can reverse the magnetization of the free layer.
- Magnetoresistance is read by passing current through the MTJ structure.
JMRAM Memory (NGC, MSU, ASU)

• Memory cell requires a special kind of magnetic Josephson junction that is switchable so that the superconducting wave function will experience either a zero or pi phase shift.

• With a pi phase shift, the magnetic junction creates a spontaneous current, and the critical current of the circuit is measurably reduced.
Why this is hard
Advanced Fabrication Process

- No commercial foundry exists that can fabricate circuits at the required level of complexity.
- MIT Lincoln Laboratory (LL) has a niobium superconductor circuit foundry that IARPA is upgrading to meet the aggressive program goals.
- With 200 mm wafers and 8+ planarized niobium layers, the MIT LL superconducting foundry is now the most advanced in the world – and continues to advance.
- LL will transfer the technology elsewhere as directed.

Test chips with up to 144 k JJs all working have been successfully demonstrated in the 8-Nb-layer process. **We are now only four orders of magnitude behind CMOS!**
Lincoln Laboratory Yield Problem

Poor cleaning after fluorine etch may lead to anodization separating from niobium electrode

- Niobium
- Anodization
- Silicon oxide
- Al - Al oxide tunnel junction

Transmission Electron Microscope Photo from Sandia

Oxygen

Fluorine
Superconducting Logic Development

- Designing complex circuits without software design tools is like *making bricks without straw*
  - Some software development is included in C3
  - SuperTools may help by C3 Phase II
- Learning requires multiple fabrication turns

- Complications include
  - Incomplete set of design rules
  - Design rules change for every advance in technology node
  - Multi-project wafer runs are inherently less reliable than single project runs
  - Unanticipated problems with fabrication

- Close communication between design teams and LL fab is essential for success
## Foundry Technology Roadmap

<table>
<thead>
<tr>
<th>Fabrication Process Attribute</th>
<th>Units</th>
<th>SFQ3ee</th>
<th>SFQ4ee</th>
<th>SFQ5ee</th>
<th>SFQ6ee</th>
<th>SFQ7ee</th>
<th>SFQ8ee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical current density</td>
<td>MA/m²</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>JJ diameter (surround)</td>
<td>nm</td>
<td>700 (500)</td>
<td>700 (500)</td>
<td>700 (300)</td>
<td>700 (300)</td>
<td>500 (200)</td>
<td>500 (200)</td>
</tr>
<tr>
<td>Nb metal layers</td>
<td>-</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Line width (space) Critical layers</td>
<td>nm</td>
<td>500 (1000)</td>
<td>500 (700)</td>
<td>350 (500)</td>
<td>350 (500)</td>
<td>250 (300)</td>
<td>180 (220)</td>
</tr>
<tr>
<td>Line width (space) Other layers</td>
<td>nm</td>
<td></td>
<td>500 (700)</td>
<td>500 (700)</td>
<td>350 (500)</td>
<td>250 (300)</td>
<td></td>
</tr>
<tr>
<td>Metal thickness</td>
<td>nm</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>150</td>
</tr>
<tr>
<td>Dielectric thickness</td>
<td>nm</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>180</td>
</tr>
<tr>
<td>Resistor width (space)</td>
<td>nm</td>
<td>1000 (2000)</td>
<td>500 (700)</td>
<td>500 (700)</td>
<td>500 (700)</td>
<td>500 (500)</td>
<td>350 (350)</td>
</tr>
<tr>
<td>Shunt resistor value</td>
<td>Ω/sq</td>
<td>2</td>
<td>2</td>
<td>2 or 6</td>
<td>2 or 6</td>
<td>2 or 6</td>
<td>2 or 6</td>
</tr>
<tr>
<td>mΩ resistor</td>
<td>mΩ</td>
<td>-</td>
<td>-</td>
<td>3 - 10</td>
<td>3 - 10</td>
<td>3 - 10</td>
<td>3 - 10</td>
</tr>
<tr>
<td>High kinetic inductance layer</td>
<td>pH/sq</td>
<td>-</td>
<td>-</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Via diameter (surround)</td>
<td>nm</td>
<td>700 (500)</td>
<td>700 (500)</td>
<td>500 (350)</td>
<td>500 (350)</td>
<td>350 (250)</td>
<td>350 (200)</td>
</tr>
<tr>
<td>Via type, stacking</td>
<td>-</td>
<td>Etched, Staggered</td>
<td>Etched, Stacked (\frac{1}{2})</td>
<td>Etched, Stacked (\frac{1}{2})</td>
<td>Etched, Stacked (\frac{1}{2})</td>
<td>Stud, Stacked</td>
<td>Stud, Stacked</td>
</tr>
<tr>
<td>Early access availability</td>
<td>-</td>
<td>2014-09</td>
<td>2015-09</td>
<td>2016-03</td>
<td>2016-09</td>
<td>2017-09</td>
<td></td>
</tr>
</tbody>
</table>

**Changes** from the previous process
Magnetic Memory Optimization

- Optimizing magnetic layers
  - Shape, size
  - Thickness
  - Materials
  - Temperature dependence
  - Fabrication
    - Crystalline anisotropy
    - Smoothness
    - Uniformity
    - Hygiene effects
- Decoders
- Drivers
- System requires integration of diverse technologies onto a single substrate
Miscellaneous
Japanese 3-year program started

- “Superconductor Electronics System Combined with Optics and Spintronics”
  JST-ALCA Project: http://www.super.nuqe.nagoya-u.ac.jp/alca/ (Japanese)
- Processor goals: AQFP logic, 8-bit simplified RISC architecture, ~25,000 JJs, ~10 instructions
Adiabatic Quantum Flux Parametron

Nobuyuki Yoshikawa, IEEE/CSC & ESAS European Superconductivity News Forum (ESNF) No. 24 April 2013

The project will be led by Professor Mark Blamire, Head of the Department of Materials Sciences at the University of Cambridge, and Dr Jason Robinson, University Lecturer in Materials Sciences, Fellow of St John’s College, University of Cambridge, and University Research Fellow of the Royal Society. They will work with partners in the University’s Cavendish Laboratory (Dr Andrew Ferguson) and at Royal Holloway, London (Professor Matthias Eschrig).

A Cambridge-led project aiming to develop a new architecture for future computing based on superconducting spintronics – technology designed to increase the energy-efficiency of high-performance computers and data storage – has been announced.

A project which aims to establish the UK as an international leader in the development of “superconducting spintronics” – technology that could significantly increase the energy-efficiency of data centres and high-performance computing – has been announced.

Led by researchers at the University of Cambridge, the “Superspin” project aims to develop prototype devices that will pave the way for a new generation of ultra-low power supercomputers, capable of processing vast amounts of data, but at a

“Superconducting spintronics offer extraordinary potential because they combine the properties of two traditionally incompatible fields to enable ultra-low power digital electronics.”

– Jason Robinson
National Strategic Computing Initiative (NSCI)

Executive Order July 29, 2015

By the authority vested in me as President by the Constitution and the laws of the United States of America, and to maximize benefits of high-performance computing research, development, and deployment, it is hereby ordered as follows:

• (b) Foundational Research and Development Agencies. There are two foundational research and development agencies for the NSCI: the Intelligence Advanced Research Projects Activity (IARPA) and the National Institute of Standards and Technology (NIST).

IARPA will focus on future computing paradigms offering an alternative to standard semiconductor computing technologies. NIST will focus on measurement science to support future computing technologies. The foundational research and development agencies will coordinate with deployment agencies to enable effective transition of research and development efforts that support the wide variety of requirements across the Federal Government.
NSCI Objectives

1. Exascale computing system (~100x performance relative to present)

2. Increase coherence between the technology base used for modeling and simulation and that used for data analytic computing.

3. Establishing, over the next 15 years, a viable path forward for future HPC systems even after the limits of current semiconductor technology are reached (the "post-Moore's Law era").

4. Increase the capacity and capability of an enduring national HPC ecosystem by employing a holistic approach that addresses relevant factors such as networking technology, workflow, downward scaling, foundational algorithms and software, accessibility, and workforce development.

5. Develop an enduring public-private collaboration to ensure that the benefits of the research and development advances are, to the greatest extent, shared between the United States Government and industrial and academic sectors.
IARPA NSCI Portfolio

• Cryogenic Computing Complexity

• Quantum Computing Programs
  – QEO: seeks tenfold increase in quantum bit coherence time
  – Multi-Qubit Coherent Operations: seeks to resolve challenges of multiple qubit technology systems
  – LogiQ: seeks to build a better logical qubit out of imperfect physical qubits

• Brain Inspired MICrONS program: seeks to revolutionize machine learning by reverse-engineering algorithms of the brain

IARPA is seeking new ideas and program managers to fill out its NSCI portfolio
Thank you!