Sirius: An Open End-to-End Voice and Vision Personal Assistant and Its Implications for Future Warehouse Scale Computers

**Johann Hauswald**, Michael A. Laurenzano, Yunqi Zhang, Cheng Li, Austin Rovinski, Arjun Khurana, Ron Dreslinski, Trevor Mudge, Vinicius Petrucci, Lingjia Tang, Jason Mars

University of Michigan — Ann Arbor, MI
Intelligent Personal Assistants (IPAs)

Apple | Google | Microsoft | Amazon
Rise of the Wearables

40%

$80bn
Scaling Current Datacenters

Ratio of IPA to Web Search Queries

0%  50%  100%
Scaling Current Datacenters

10% IPA: 16x Machines

Ratio of IPA to Web Search Queries

Compute Resources
Scaling Current Datacenters

![Diagram showing the scaling of datacenters with different ratios of IPA to Web search queries and corresponding 50% IPA: 80x Machines.]
Scaling Current Datacenters

- 100% IPA: 160x Machines

Ratio of IPA to Web Search Queries

Compute Resources

100

10

1

0% 50% 100%
The Challenge

Redesign the datacenter for intelligent personal assistants

No Open Source IPA

Investigate Future Datacenters Designs

Sirius: An Open End-to-End Voice and Vision Personal Assistant
Open Source
Intelligent Personal Assistant

Benchmark Suite

Ported Suite Across
Accelerator Platforms

Investigate Future Datacenter Designs For IPAs
Sirius
Sirius: An Open End-to-End Voice and Vision Personal Assistant
Sirius: An Open End-to-End Voice and Vision Personal Assistant
“Set my alarm for 6am”
“What is the capital of Turkey?”
“What is the capital of Turkey?”
“Ankara”

Users

Display Answer

Image

Image Matching

Image Database

Question-Answering

Search Database

Mobile Server

Automatic Speech-Recognition

Query Classifier

Question or Action

Display Answer

“Ankara”

Users

Display Answer
Sirius: An Open End-to-End Voice and Vision Personal Assistant

Users

Image

Display Answer

Image Matching

Image Database

Question-Answering

Search Database

Voice

Automatic Speech-Recognition

Query Classifier

Question or Action

Mobile Server

Execute Action

Answer

Database

Image Data

Question

Action

“What is the capital of Turkey?”
“How tall is the eiffel tower?”

Users → Image → Mobile Server → Automatic Speech-Recognition

 Execute Action

Display Answer

Image Database

Image Matching

Question-Answering

Search Database

“How tall is the eiffel tower?”
Sirius: An Open End-to-End Voice and Vision Personal Assistant

Users

Display Answer

Image

Mobile Server

Voice

Execute Action

Image Matching

Automatic Speech-Recognition

Query Classifier

Question or Action

Image Database

Image Data

Answer

Question-Answering

Search Database

“How tall is the Eiffel Tower?”

“How is the capital of Turkey?”
“300 meters”
- **Sirius**: full end-to-end with inputs, pre-trained models, and databases
- **Sirius-suite**: 7 kernels with inputs to study each service individually

{sirius.clarity-lab.org}
- **Sirius**: full end-to-end with inputs, pre-trained models, and databases
- **Sirius-suite**: 7 kernels with inputs to study each service individually

[ sirius.clarity-lab.org ]
How does Sirius work?

**Query Taxonomy**
- Voice Command (VC)
- Voice Query (VQ)
- Voice-Image Query (VIQ)

**IPA Services**
- Automatic-Speech Recognition (ASR)
- Question Answering (QA)
- Image Matching (IMM)

**Tasks**
- Signal Processing
- Natural Language Processing
- Image Processing

**Open Source Tools**
- CMU Sphinx
- Ephyra
- KALDI
- OpenCV
Sirius-suite
Sirius-suite

Automatic-Speech Recognition (ASR)

Question Answering (QA)

Image Matching (IMM)

IPA Services
Sirius-suite

IPA Services

Automatic-Speech Recognition (ASR)

Question Answering (QA)

Image Matching (IMM)

Gaussian Mixture Model

Deep Neural Network

HMM

85%
15%
78%
22%
Sirius-suite

Automatic-Speech Recognition (ASR)

- GMM (85%)
- DNN (78%)

Question Answering (QA)

Image Matching (IMM)

IPA Services
Sirius-suite

Automatic-Speech Recognition (ASR)
- GMM (85%)
- DNN (78%)

Question Answering (QA)

Image Matching (IMM)

IPA Services

- GMM (85%)
- DNN (78%)

Conditional Random Fields

Regex

Search

Other

46%

12%

17%

3%

22%
Sirius-suite

Automatic-Speech Recognition (ASR)
- GMM (85%)
- DNN (78%)

Question Answering (QA)
- Stemmer (46%)
- Regex (22%)
- CRF (17%)

Image Matching (IMM)

IPA Services
Sirius-suite

Automatic-Speech Recognition (ASR)

- GMM (85%)
- DNN (78%)

Question Answering (QA)

- Stemmer (46%)
- Regex (22%)
- CRF (17%)

Image Matching (IMM)

IPA Services

Feature Extraction

- ANN: 3%
- 41%

Feature Description

- 56%
Sirius-suite

Automatic-Speech Recognition (ASR)
- GMM (85%)
- DNN (78%)

Question Answering (QA)
- Stemmer (46%)
- Regex (22%)
- CRF (17%)

Image Matching (IMM)
- FE (41%)
- FD (56%)

IPA Services
Sirius-suite

Automatic-Speech Recognition (ASR)
- GMM (85%)
- DNN (78%)

Question Answering (QA)
- Stemmer (46%)
- Regex (22%)
- CRF (17%)

Image Matching (IMM)
- FE (41%)
- FD (56%)

IPA Services

7 kernels: 92% total execution of Sirius
Suite entirely written in C/C++/CUDA
Release includes inputs and models
Future Datacenter Design
How must current datacenters be upgraded to meet demand?

What is the efficiency of the upgraded datacenter?
Upgrading Datacenters with COTS Systems

<table>
<thead>
<tr>
<th>Platform</th>
<th>Model</th>
<th>Clock</th>
<th>Threads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multicore CPU</td>
<td>Intel Xeon E3-1240 V3</td>
<td>3.40 GHz</td>
<td>8</td>
</tr>
<tr>
<td>GPU</td>
<td>NVIDIA GTX 770</td>
<td>1.05 GHz</td>
<td>12288</td>
</tr>
<tr>
<td>Intel Phi</td>
<td>Phi 5110P</td>
<td>1.05 GHz</td>
<td>240</td>
</tr>
<tr>
<td>FPGA</td>
<td>Xilinx Virtex-6 ML605</td>
<td>400 MHz</td>
<td>N/A</td>
</tr>
</tbody>
</table>
## Upgrading Datacenters with COTS Systems

<table>
<thead>
<tr>
<th>Platform</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multicore CPU</td>
<td>Minor SW changes</td>
<td>Limited speedup</td>
</tr>
<tr>
<td>GPU</td>
<td>Many threads</td>
<td>Programability</td>
</tr>
<tr>
<td>Intel Phi</td>
<td>Manycore</td>
<td>Limited compiler support</td>
</tr>
<tr>
<td>FPGA</td>
<td>Flexible</td>
<td>New implementation</td>
</tr>
</tbody>
</table>
## Acceleration Overview

<table>
<thead>
<tr>
<th>Platform</th>
<th>GMM</th>
<th>DNN</th>
<th>Stemmer</th>
<th>Regex</th>
<th>CRF</th>
<th>FE</th>
<th>FD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMP</td>
<td>3.5</td>
<td>6.0</td>
<td>4.0</td>
<td>3.9</td>
<td>3.7</td>
<td>5.2</td>
<td>5.9</td>
</tr>
<tr>
<td>GPU</td>
<td>70.0</td>
<td>54.7</td>
<td>6.2</td>
<td>48.0*</td>
<td>3.8*</td>
<td>10.5</td>
<td>120.5</td>
</tr>
<tr>
<td>Intel Phi</td>
<td>1.1</td>
<td>11.2</td>
<td>5.6</td>
<td>1.1</td>
<td>4.7</td>
<td>2.5</td>
<td>12.7</td>
</tr>
<tr>
<td>FPGA</td>
<td>169.0</td>
<td>110.5*</td>
<td>30.0</td>
<td>168.2*</td>
<td>7.5*</td>
<td>34.6*</td>
<td>75.5*</td>
</tr>
</tbody>
</table>
# Acceleration Overview

<table>
<thead>
<tr>
<th>Platform</th>
<th>GMM</th>
<th>DNN</th>
<th>Stemmer</th>
<th>Regex</th>
<th>CRF</th>
<th>FE</th>
<th>FD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMP</td>
<td>3.5</td>
<td>6.0</td>
<td>4.0</td>
<td>3.9</td>
<td>3.7</td>
<td>5.2</td>
<td>5.9</td>
</tr>
<tr>
<td>GPU</td>
<td>70.0</td>
<td>54.7</td>
<td>5.2</td>
<td>18.0</td>
<td>3.9</td>
<td>10.5</td>
<td>120.5</td>
</tr>
<tr>
<td>Intel Phi</td>
<td>11.2</td>
<td>5.6</td>
<td>1.1</td>
<td>4.7</td>
<td>2.5</td>
<td>12.7</td>
<td></td>
</tr>
<tr>
<td>FPGA</td>
<td>169.0</td>
<td>110.5*</td>
<td>30.0</td>
<td>168.2*</td>
<td>7.5*</td>
<td>34.6*</td>
<td>75.5*</td>
</tr>
</tbody>
</table>

*Custom Porting: 75% of the Implementations*
Acceleration Results

![Speedup Heatmap](image)

Platform:
- FPGA
- Phi
- GPU
- CMP

Workload:
- GMM
- DNN
- Stemmer
- Regex
- CRF
- FE
- FD

Speedup:
- 20
Acceleration Results

![Speedup Heatmap](Image)

- **Platform**:
  - FPGA
  - Phi
  - GPU
  - CMP

- **Workload**:
  - GMM
  - DNN
  - Stemmer
  - Regex
  - CRF
  - FE
  - FD

- **Speedup**

---

Sirius: An Open End-to-End Voice and Vision Personal Assistant
Acceleration Results

~6x

FPGA

~5x

Speech Recognition

Question Answering

Image Matching

Speedup

GMM  DNN  StemmerRegex  CRF  FE  FD

Workload Speedup Heatmap
Acceleration Results

- **Speedup**
  - Image Matching: 120x
  - Question Answering: ~52x
  - Speech Recognition: ~52x

**Platform**
- FPGA
- Phi
- GPU
- CMP

**Workload Speedup Heatmap**
- GMM
- DNN
- StemmerRegex
- CRF
- FE
- FD
Acceleration Results

Sirius: An Open End-to-End Voice and Vision Personal Assistant
Service Latency Improvement

- Single Core
- CMP (sub-query)
- GPU: 0.12
- Phi
- FPGA: 0.05
Service Latency Improvement

Latency (s)

Platform

- Single Core
- CMP (sub-query)
- GPU: 0.12
- Phi
- FPGA: 0.05

2.8s
Service Latency Improvement

Using 8 threads

Platform

- Single Core
- CMP (sub-query)
- GPU
- Phi
- FPGA

Latency (s)

- IMM_FE
- IMM_FD

Using 8 threads
Service Latency Improvement

Latency (s) (* includes DNN and HMM combined)

ASR_GMM  ASR_HMM  ASR_DNN  ASR_HMM

Single Core
CMP (sub-query)
GPU  0.22
Phi
FPGA  0.19

IMM_FE  IMM_FD

Single Core
CMP (sub-query)
GPU  0.12
Phi
FPGA  0.05

QA_Stemmer  QA_CRF  QA_Regex  QA_Other

0.94
Service Latency Improvement

Average Latency Reduction:
FPGA: 16x
GPU: 10x
Performance improvements increase throughput

Reduce the number of servers
Performance improvements increase throughput

Reduce the number of servers

What is the Total Cost of Ownership of an accelerator upgraded Datacenter?
### TCO Model Parameters [1]

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server Price</td>
<td>$2,102</td>
</tr>
<tr>
<td>Server Power</td>
<td>164 W</td>
</tr>
<tr>
<td>PUE</td>
<td>1.1</td>
</tr>
<tr>
<td>DC Depreciation</td>
<td>12 years</td>
</tr>
<tr>
<td>Server Depreciation</td>
<td>3 years</td>
</tr>
<tr>
<td>Average Server Utilization</td>
<td>45%</td>
</tr>
<tr>
<td>Electricity Cost</td>
<td>0.067/kWh</td>
</tr>
<tr>
<td>Datacenter Price</td>
<td>$10/W</td>
</tr>
<tr>
<td>Datacenter Opex</td>
<td>$0.04/W</td>
</tr>
<tr>
<td>Server Opex</td>
<td>5% of Capex/year</td>
</tr>
</tbody>
</table>

TCO — Query Level Results

Improvement

TCO

3x
2.5x
2x
1.5x
1x

GPU
FPGA
TCO — Query Level Results

Average TCO improvement:
- GPU: 2.6x
- FPGA: 1.4x
Other topics included in the paper:

- Real System Analysis
- Question Variability Analysis
- Accelerator Performance per Watt
- Throughput Improvement at Various Load Levels
- Accelerator Porting Methodology
- FPGA Implementation Details
- Homogeneous/Heterogeneous Datacenter Design
**Sirius**: full application

**Sirius-suite**: 7 kernels to study each service

[sirius.clarity-lab.org](http://sirius.clarity-lab.org)
Thank you