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, Ronald G. Dreslinski, Trevor Mudge, Vinicius Petrucci[1], Lingjia Tang, Jason Mars

Clarity Lab
University of Michigan - Ann Arbor, MI, USA
{jahausw, mlaurenz, yunqi, elfchris, rovinski, khuranaa, rdreslin, tnm, lingjia, profmars}@umich.edu

Abstract

As user demand scales for *intelligent personal assistants* (IPAs) such as Apple's Siri, Google's Google Now, and Microsoft's Cortana, we are approaching the computational limits of current datacenter architectures. It is an open question how future server architectures should evolve to enable this emerging class of applications, and the lack of an open-source IPA workload is an obstacle in addressing this question.

In this paper, we present the design of **Sirius**, an open end-to-end IPA web-service application that accepts queries in the form of voice and images, and responds with natural language. We then use this workload to investigate the implications of four points in the design space of future accelerator-based server architectures spanning traditional CPUs, GPUs, manycore throughput co-processors, and FPGAs. To investigate future server designs for Sirius, we decompose Sirius into a suite of 7 benchmarks (**Sirius Suite**) comprising the computationally intensive bottlenecks of Sirius. We port Sirius Suite to a spectrum of accelerator platforms and use the performance and power trade-offs across these platforms to perform a total cost of ownership (TCO) analysis of various server design points. In our study, we find that accelerators are critical for the future scalability of IPA services. Our results show that GPU- and FPGA-accelerated servers improve the query latency on average by 10× and 16×. For a given throughput, GPU- and FPGA-accelerated servers can reduce the TCO of datacenters by 2.6× and 1.4×, respectively.

ASPLOS '15, March 14–18, 2015, Istanbul, Turkey.

http://dx.doi.org/10.1145/2694344.2694347

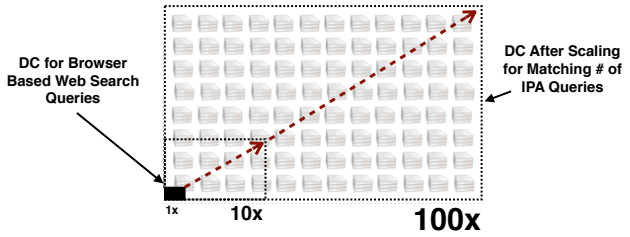


Figure 1: Impact of Higher Computational Requirements for IPA Queries on Datacenters (DCs)

Categories and Subject Descriptors C.0 [*Computer Systems Organization*]: General—System architectures; C.5.5 [*Computer System Implementation*]: Servers

Keywords datacenters; warehouse scale computers; emerging workloads; intelligent personal assistants

1. Introduction

Apple's Siri [1], Google's Google Now [2] and Microsoft's Cortana [3] represent a class of emerging web-service applications known as *Intelligent Personal Assistants* (IPAs). An IPA is an application that uses inputs such as the user's voice, vision (images), and contextual information to provide assistance by answering questions in natural language, making recommendations, and performing actions. These IPAs are emerging as one of the fastest growing Internet services as they have recently been deployed on well known platforms such as iOS, Android, and Windows Phone, making them ubiquitous on mobile devices worldwide [4]. In addition, the usage scenarios for IPAs are rapidly increasing with recent offerings in wearable technologies such as smart watches [5] and smart glasses [6]. Recent projections predict the wearables market to be at 485 million annual device shipments by 2018 [7]. This growth in market share, coupled with the fact that the design of wearables is heavily reliant on voice

[1] Work conducted as a postdoctoral fellow in Clarity Lab at the University of Michigan. He is now a professor at Federal University of Bahia.

and image input, further indicates that rapid growth in user demand for IPA services is on the horizon.

IPAs differ from many of the web-service workloads currently present in modern warehouse-scale computers (WSCs). In contrast to the queries of traditional browser-centric services, IPA queries stream through software components that leverage recent advances in speech recognition, natural language processing and computer vision to provide users a speech-driven and/or image-driven contextually-based question-and-answer system [8]. Due to the computational intensity of these components and the large data-driven models they use, service providers house the required computation in massive datacenter platforms in lieu of performing the computation on the mobile devices themselves. This offloading approach is used by both Apple's Siri and Google's Google Now as they send compressed recordings of voice command/queries to datacenters for speech recognition and semantic extraction [9]. However, datacenters have been designed and tuned for traditional web services such as Web Search and questions arise as to whether the current design employed by modern datacenters, composed of general-purpose servers, is suitable for emerging IPA workloads.

IPA queries require a significant amount of compute resources compared to traditional text-based web services such as Web Search. As we show later in this work, the computational resources required for a single leaf query is in excess of 100× more than that of traditional Web Search. Figure 1 illustrates the scaling of compute resources in a modern datacenter required to sustain an equivalent throughput of IPA queries compared to Web Search. Due to the looming *Scalability Gap* shown in the figure, there has been significant interest in both academia and industry to leverage hardware acceleration in datacenters using various platforms such as GPU, manycore co-processors and FPGAs to achieve high performance and energy efficiency. To gain further insight on whether there are sufficient acceleration opportunities for IPA workloads and what the best acceleration platform is, several challenges need to be addressed, including:

1. Identifying critical compute and performance bottlenecks throughout the end-to-end lifetime of an IPA query;
2. Understanding the performance, energy and cost trade-offs among popular accelerator options given the characteristics of IPA workloads;
3. Designing future server and datacenter solutions that can meet the amount of future user demand while being cost and energy efficient.

However, the lack of a representative, publicly available, end-to-end IPA system proves prohibitive for investigating the design space of future accelerator-based server designs for this emerging workload. To address this challenge, we first construct an end-to-end standalone IPA service, **Sirius**, that implements the core functionalities of an IPA such as speech recognition, image matching, natural language processing and a question-and-answer system. Sirius takes as input user dictated speech and/or image(s) captured by a camera. There are three pathways of varying complexity through the Sirius back-end based on the nature of the input query. A voice command primarily exercises speech recognition on the server-side to execute a command on the mobile device. A voice query additionally leverages a sophisticated natural language processing (NLP) question-and-answer system to produce a natural language response to the user. A voice and image question such as *"When does this restaurant close?"* coupled with an image of the restaurant, also leverages image matching with an image database and combines the matching output with the voice query to select the best answer for the user. We have constructed Sirius by integrating three services built using well-established open source projects that include techniques and algorithms representative of those found in commercial systems. These open projects include CMU's Sphinx [10], representing the widely-used Gaussian Mixture Model based speech recognition, Kaldi [11] and RWTH's RASR [12], representing industry's recent trend toward Deep Neutral Network based speech recognition, OpenEphyra [13] representing the-state-of-the-art question-and-answer system based on IBM's Watson [14], and SURF [15] implemented using OpenCV [16] representing state-of-the-art image matching algorithms widely used in various production applications.

With this end-to-end workload in hand, we perform an in-depth investigation of the viability of various acceleration strategies, and provide insights on future datacenter and server designs for this emerging workload. Specifically, this paper makes the following contributions:

- **Sirius** - We construct Sirius, an open end-to-end intelligent personal assistant system with both speech and image front-ends. In addition to Sirius itself, we compile a query taxonomy spanning three classes of queries: Voice Command, Voice Query, and Voice/Image Query. (Section 2)
- **Scalability Gap** - We characterize Sirius on commodity hardware and demonstrate the *Scalability Gap* for this type of workload. We observe that the compute resources needed to sustain this workload is orders of magnitude higher than traditional datacenter workloads. We also perform an analysis of the cycle breakdown of IPA queries and analyze the computational bottlenecks of Sirius. We show that there is a limited speedup potential for this workload on general-purpose processors and acceleration is indeed needed to address the scalability gap. (Section 3)
- **Accelerating Sirius** - Based on our cycle breakdown analysis, we extract 7 computational bottlenecks comprising 92% of the cycles consumed by Sirius to compose a C/C++ benchmark suite (**Sirius Suite**) for acceleration. We port these workloads and conduct a thorough performance evaluation on a spectrum of accelerator platforms. The end-to-end Sirius, query taxonomy, input set, Sirius Suite benchmarks, and the full source code ported across accelerators are available online [17]. (Section 4)

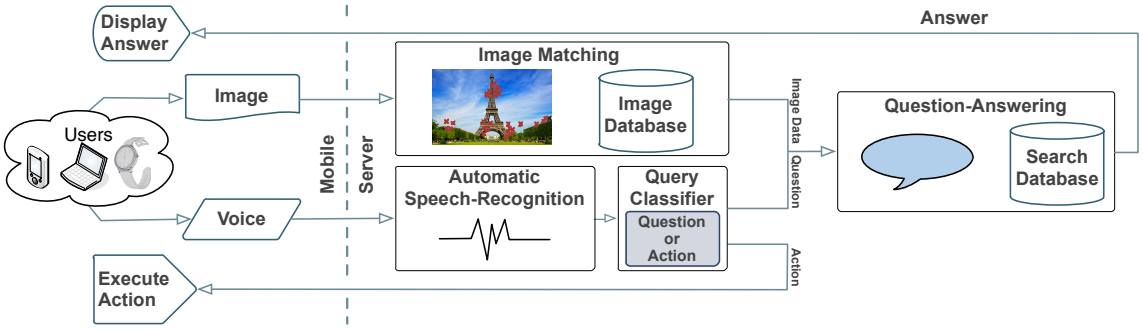


Figure 2: End-to-end Diagram of the Sirius Pipeline

- **Future Server and Datacenter Design** - Based on our acceleration results, we investigate the implications for future server designs. After evaluating the trade-offs between performance, power efficiency and the total cost of ownership of a datacenter, we propose server and datacenter designs that significantly reduce the computation gap between user demand and the current datacenter's computation capability. (Section 5)

In summary, we find that among the popular acceleration options including GPU, Intel Phi and FPGA, the FPGA-accelerated server is the best server option for a homogeneous datacenter design when the design objective is minimizing latency or maximizing energy efficiency with a latency constraint. FPGA achieves an average 16× reduction on the query latency across various query types over the baseline multicore system. On the other hand, GPUs provide the highest TCO reduction on average. GPU-accelerated servers can achieve an average 10× query latency reduction, translating to a 2.6× TCO reduction. When excluding FPGAs as an acceleration option, GPUs provide the best latency and cost reduction among the rest of the accelerator choices. On average, replacing FPGAs using GPUs leads to a 66% longer latency, but in return achieves a 47% TCO reduction and simpler software engineering costs.

2. Sirius: An End-to-End IPA

In this section we present **Sirius**: an end-to-end intelligent personal assistant (IPA). We first describe the design objectives for Sirius, then present an overview of Sirius and a taxonomy of query types it supports. Finally, we detail the underlying algorithms and techniques used by Sirius.

2.1 Sirius Design Objectives

There are three key objectives in the design for Sirius:

1. **[Completeness]** - Sirius should provide a complete IPA service that takes the input of human voice and images and provide a response to the user's question with natural language.
2. **[Representativeness]** - The computational techniques used by Sirius to provide this response should be representative of state-of-the-art approaches used in commercial domains.
3. **[Deployability]** - Sirius should be deployable and fully functional on real systems.

2.2 Sirius Overview: Life of an IPA Query

Figure 2 presents a high-level diagram of the end-to-end Sirius query pipeline. The life of a query begins with a user's voice and/or image input through a mobile device. Compressed versions of the voice recording and image(s) are sent to a server housing Sirius. The user's voice is then processed by an Automatic Speech Recognition (ASR) front-end that translates the user's speech question into its text equivalent using statistical models. The translated speech then goes through a Query Classifier (QC) that decides if the speech is an action or a question. If it is an action, the command is sent back to the mobile device for execution. Otherwise, the Sirius back-end receives the question in plain text. Using natural language processing (NLP) techniques, the Question- Answering (QA) service extracts information from the input, searches its database, and chooses the best answer to return to the user. If an image accompanies the speech input, Sirius uses computer vision techniques to attempt to match the input image to its image database and return relevant information about the matched image using the Image Matching (IMM) service. For example, a user can ask *"What time does this restaurant close?"* while image(s) of the restaurant are captured via smart glasses [6]. Sirius can then return an answer to the query based not only on the speech, but also information from the image.

As shown in Figure 2, there are a number of pathways a single query can take based on the type of directive, whether it be question or action, and the type of input, speech only or accompanied by images. In order to design the input set used with Sirius we have identified a query taxonomy of three classes that covers these pathways. Table 1 summarizes these query classes providing an example for each, the Sirius services they exercise, the resulting behavior of Sirius, and the number of queries of that type in our input set. Figure 3 illustrates a tiered view of Sirius spanning the query taxonomy it supports, the services that comprise Sirius, and the algorithmic sub-components that compose each service. We describe these services and algorithms in the following section.

Table 1: Query Taxonomy

Query Type	Example	Service	Result	# of Queries
Voice Command (VC)	"Set my alarm for 8am."	ASR	Action on user's device	16
Voice Query (VQ)	"Who was elected 44th president?"	ASR & QA	Best answer from QA	16
Voice-Image Query (VIQ)	"When does this restaurant close?"	ASR, QA & IMM	Best results from IMM and QA	10

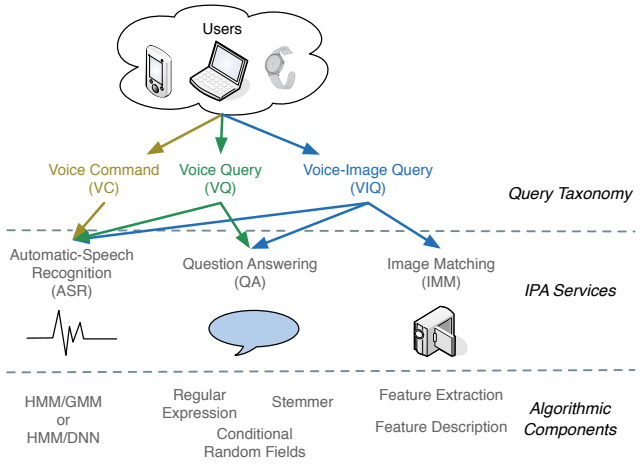


Figure 3: Tier-level View of Sirius

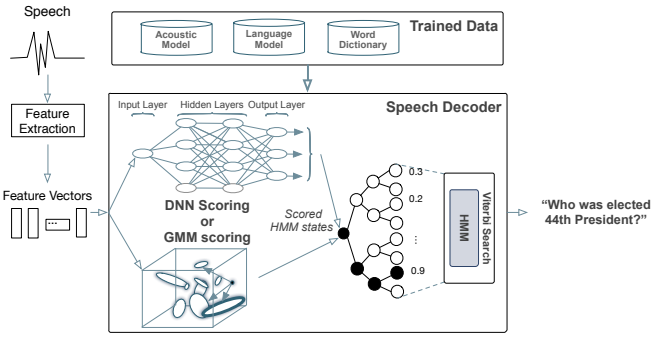


Figure 4: Automatic Speech Recognition Pipeline

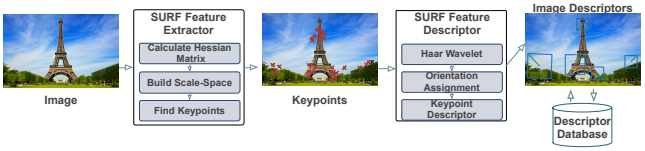


Figure 5: Image Matching Pipeline

2.3 The Design of Sirius: IPA Services and Algorithmic Components

As shown in Figure 3, Sirius is composed of three IPA services: speech recognition (ASR), question-answering (QA), and image matching (IMM). These services can be further decoupled into their individual algorithmic components. In order to design Sirius to be representative of production grade systems, we leverage well-known open infrastructures that use the same algorithms as commercial applications. Speech recognition in Google Voice, for example, has used speaker-independent Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) and is adopting Deep Neural Networks (DNNs) [18, 19]. The OpenEphyra framework used for question-answering is an open-source release from CMU's prior research collaboration with IBM on the Watson system [14]. OpenEphyra's NLP techniques, including conditional random field (CRF), have been recognized as state-of-the-art and are used at Google and in other industry question-answering systems [20]. We design our image matching pipeline based on the SURF algorithm, which is widely used in industry [15, 21, 22]. We implement SURF using the open source computer vision (OpenCV) library [16], which is employed in commercial products from companies like Google, IBM, and Microsoft. The design of these services are described in the remainder of this section.

2.3.1 Automatic-Speech Recognition (ASR)

The inputs to the ASR are feature vectors representing the speech segment, generated by fast pre-processing and feature extraction of the speech. The ASR component relies on a combination of a Hidden Markov Model (HMM) and either a Gaussian Mixture Model (GMM) or a Deep Neural Network (DNN). Sirius' GMM-based ASR uses CMU's Sphinx [10], while the DNN-based ASR includes Kaldi [11] and RWTH's RASR [12].

As shown in Figure 4, the HMM builds a tree of states for the current speech frame using input feature vectors. The GMM or DNN scores the probability of the state transitions in the tree, and the Viterbi algorithm [23] then searches for the most likely path based on these scores. The path with the highest probability represents the final translated text output. The GMM scores HMM state transitions by mapping an input feature vector into a multi-dimensional coordinate system and iteratively scores the features against the trained acoustic model.

DNN, however, scores using probabilities from a neural network. The depth of a DNN is defined by the number of hidden layers where scoring amounts to one forward pass through the network. In recent years, industry and academia have moved towards DNN over GMM due to its higher accuracy [24, 25].

2.3.2 Image Matching (IMM)

The image matching pipeline receives an input image, attempts to match it against images in a pre-processed image database, and returns information about the matched

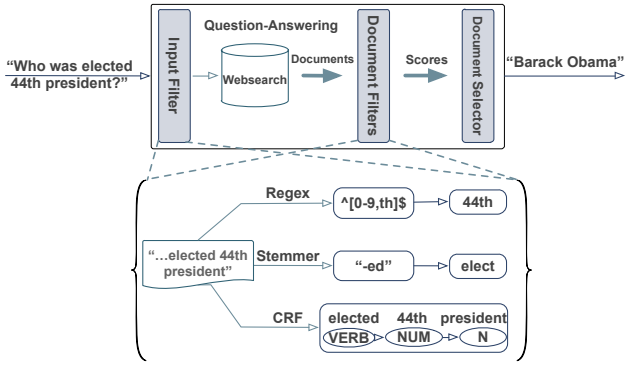
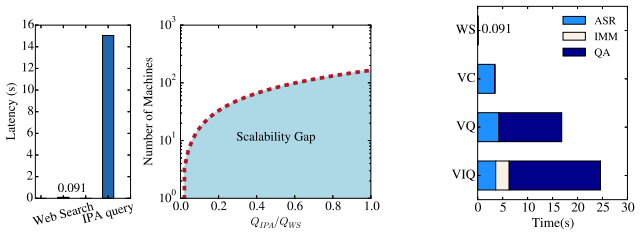


Figure 6: OpenEphyra Question-Answering Pipeline



(a) Scalability Gap

(b) Latency Across Query Types

Figure 7: Scalability Gap and Latency

images. The database that is used in Sirius is the Mobile Visual Search [26] database. Image keypoints are first extracted from the input image using the SURF algorithm [15]. In Feature Extraction (FE), the image is downsampled and convolved multiple times to find interesting points at different scales. After thresholding the convolution responses, the local maxima responses are stored as image keypoints. Figure 5 details the steps in this process. The keypoints are then passed to the Feature Descriptor (FD) component where they are assigned an orientation vector, and similarly oriented keypoints are grouped into feature descriptors. This process reduces variability across input images, increasing chances of finding the correct match. The descriptors from the input image are matched to pre-clustered descriptors representing the database images by using an approximate nearest neighbor (ANN) search; the database image with the highest number of matches is returned.

2.3.3 Question-Answering (QA)

The text output from the ASR is passed to OpenEphyra (OE) [13], which uses three core processes to extract textual information: word stemming, regular expression matching, and part-of-speech tagging. Figure 6 shows a diagram of the OE engine incorporating these components, generating Web Search queries and filtering the returned results. The Porter Stemming [27] algorithm (stemmer) exposes the root of a word by matching and truncating common word endings. OE also uses a suite of regular-expression patterns to match common query words (what, where, etc) and filter any special characters in the input. The Conditional Random Field (CRF) classifier [28] takes a sentence, the position of each word in the sentence, and the label of the current and previous word as input to makes predictions on the part-of-speech for each word of an input query. Each input query is parsed using the aforementioned components to generate queries to the web search engine. Next, filters using the same techniques are used to extract information from the returned documents; the document with the highest overall score after score aggregation is returned as the best answer.

3. Real System Analysis for Sirius

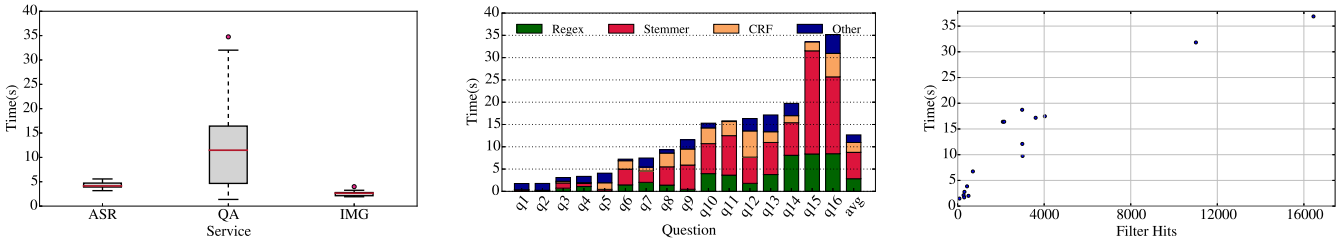
In this section, we present a real-system analysis of Sirius. The experiments throughout this section are performed using an Intel Haswell server (details in Table 3).

Scalability Gap - To gain insights on the required resource scaling for IPA queries in modern datacenters, we juxtapose the computational demand of an average Sirius query with that of an average Web Search query. To perform this experiment, we compare the average query latency (execution time) for both applications on a single core at a very low load. Both Sirius and Web Search are configured to be memory resident and go no further than main memory to process a query (i.e., minimum I/O activities).

Figure 7a (left) presents the average latency of both Web Search using open source Apache Nutch [29, 30] and Sirius queries. As shown in the figure, the average Nutch-based Web Search query latency is 91ms on the Haswell based server. In contrast, Sirius query latency is significantly longer, averaging around 15s across 42 queries spanning our three query classes (VC, VQ and VIQ from Table 1). Based on this significant difference in the computational demand, we perform a back-of-the-envelope calculation of how the compute resources (machines) in current datacenters must scale to match the throughput in queries for IPAs and Web Search.

Figure 7a (right) presents the number of machines needed to support IPA queries as the number of these queries increases. The x-axis shows the ratio between IPA queries and traditional Web Search queries. The y-axis shows the ratio of compute resources needed to support IPA queries relative to Web Search queries. As shown in the figure, current datacenter infrastructures will need to scale its compute resources to 165× its current size when the number of IPA queries scale to match the number of Web Search queries. We refer to this throughput difference as the *scalability gap*.

Sirius Query Deep Dive - To better understand the IPA query characteristics, we further investigate the average latency and latency distributions of various query types for Sirius. Figure 7b presents the average latency across query types including traditional Web Search (WS), Voice Command (VC), Voice Query (VQ) and Voice Image Query (VIQ). As shown in the figure, the latency of all three Sirius query types are significantly higher than that of Web



(a) Latency Variability Across Services (b) OpenEphyra Breakdown (c) Latency and Filter Hits in OpenEphyra

Figure 8: Sirius Variability Across Query Types and Causes

Search queries. The shortest query type is VC, which only uses the ASR service. Yet it still requires orders of magnitude more computation than Web Search. The longest query type is VIQ, which uses all three services including ASR, IMM, and QA. Among all three services, QA consistently consumes the most compute cycles.

Table 2: Voice Query Input Set

Q#	Query
q1	"Where is Las Vegas?"
q2	"What is the capital of Italy?"
q3	"Who is the author of Harry Potter?"
...	...
q15	"What is the capital of Cuba?"
q16	"Who is the current president of the United States?"

Figure 8a presents the latency distribution for each Sirius service. As shown in the figure, QA has the highest variability in latency, ranging from 1.7s to 35s depending on the input query. Figure 8b further presents the breakdown of execution time among QA's hot components (described later in this section) across the complete VQ query input set (shown in Table 2). The reason for this high latency variability is not immediately clear from inspecting the query input set, especially when considering the small difference between Q2 and Q15 in Table 2. However, after further investigation, we identified that the high variance is primarily due to the runtime variability of various document filters in the NLP component used to select the most fitting answer for a given query. Figure 8c demonstrates the correlation between latency and the number of hits in the document filters. The other services, ASR and IMM, have very low query to query variability. Next, we investigate the cycle breakdown of the algorithmic components that comprise each service.

Cycle Breakdown of Sirius Services - To identify the computational bottlenecks of each service, we perform top-down profiling of hot algorithmic components for each service, shown in Figure 3, using Intel VTune [31]. Figure 9 presents the average cycle breakdown results. Across services, a few hot components emerge as good candidates for acceleration. For example, a high percentage of the execution for ASR is spent on scoring using either GMM or DNN. For QA, on average 85% of the cycles are spent in three components including stemming, regular expression pattern matching and CRF, and for IMM, the majority of cycles are

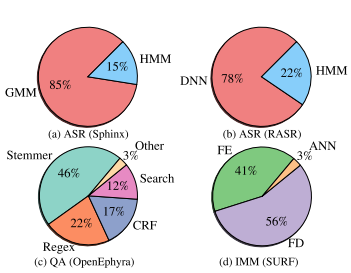


Figure 9: Cycle Breakdown per Service

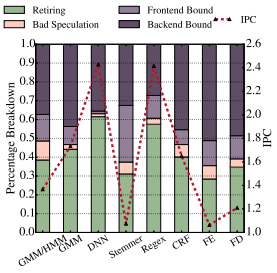


Figure 10: IPC and Bottleneck Breakdown

spent either performing feature extraction or description using the SURF algorithm.

We then identify the architectural bottlenecks for these hot components to investigate the performance improvement potential for a general-purpose processor. Figure 10 presents the instructions per cycle (IPC) and potential architectural bottlenecks (including front-end, speculation and back-end) for each component, identified using Intel VTune [31]. A few of the service components including DNN and Regex execute relatively efficiently on Xeon cores. This graph indicates that even with all stall cycles removed (i.e., perfect branch prediction, infinite cache, etc) the maximum speedup is bound by around 3×. Considering the orders of magnitude difference indicated by the scalability gap, further acceleration is needed to bridge the gap.

4. Accelerating Sirius

In this section, we describe the platforms and methodology used to accelerate the key components of Sirius. We also present and discuss the results of accelerating each of these components across 4 different accelerator platforms.

4.1 Accelerator Platforms

We use a total of four platforms, summarized in Table 3, to accelerate Sirius. Our baseline platform is an Intel Xeon Haswell CPU running single-threaded kernels.

We summarize the advantages and disadvantages of each accelerator platform below.

Table 3: Platform Specifications

Model	Multicore Intel Xeon E3-1240 V3	GPU NVIDIA GTX 770	Phi Intel Xeon Phi 5110P	FPGA Xilinx Virtex-6 ML605
Frequency	3.40 GHz	1.05 GHz	1.05 GHz	400 MHz
# Cores	4	8*	60	N/A
# HW Threads	8	12288	240	N/A
Memory	12 GB	2 GB	8 GB	512 MB
Memory BW	25.6 GB/s	224 GB/s	320 GB/s	6.40 GB/s
Peak TFLOPS	0.5	3.2	2.1	0.5

** Core = SM (Streaming Multiprocessor), 2048 threads/SM*

Table 4: Sirius Suite and Granularity of Parallelism

Service	Benchmark	Baseline	Input Set	Data Granularity
ASR	*Gaussian Mixture Model (GMM)*	CMU Sphinx [10]	HMM states	For each HMM state
	Deep Neural Network (DNN)	RWTH RASR [12]	HMM states	For each matrix multiplication
QA	*Porter Stemming (Stemmer)*	Porter [27]	4M word list	For each individual word
	Regular-Expression (Regex)	SLRE [32]	100 expressions/400 sentences	For each regex-sentence pair
	Conditional Random Fields (CRF)	CRFsuite [33]	CoNLL-2000 Shared Task [34]	For each sentence
IMM	*Feature Extraction (FE)*	SURF [15]	JPEG Image	For each image tile
	Feature Description (FD)	SURF [15]	Vector of Keypoints	For each keypoint

- **Multicore CPU** - *Advantages:* High clock frequency, not limited by branch divergence. *Disadvantages:* Least amount of threads available.
- **GPU** - *Advantages:* Massively parallel. *Disadvantages:* Power hungry, custom ISA, hard to program, large data transfer overheads, limited branch divergence handling.
- **Intel Phi** - *Advantages:* Many core, standard programming model (same ISA), manual porting optional / compiler help, handles branch divergence, high bandwidth. *Disadvantages:* Data transfer overheads, relies on compiler. *Note:* 1 core is used for the operating system running on the device itself.
- **FPGA** - *Advantages:* Can be tailored to implement very efficient computation and data layout for the workload. *Disadvantages:* Runs at a much lower clock frequency, expensive, hard to develop for and maintain with software updates.

4.2 Sirius Suite: A Collection of IPA Compute Bottlenecks

To investigate the viability and trade-offs of accelerating IPAs, we extract the key computational bottlenecks of Sirius (described in Section 3) to construct a suite of benchmarks we call **Sirius Suite**. Sirius Suite as well as its implementations across the described accelerator platforms are available alongside the end-to-end Sirius application [17]. As a basis for Sirius Suite, we port existing open-source C/C++ implementations available for each algorithmic component to our target platforms. We additionally implemented standalone C/C++ benchmarks based on the source code of Sirius where none were currently available. The baseline implementations are summarized in column 2 of Table 4. For each Sirius Suite benchmark, we built an input set representative of IPA queries. Table 4 shows the granularity at which each thread performs the computation on the accelerators. For example, both GMM and DNN kernels receive input feature vectors from the HMM search, which are all scored in parallel but at different levels of abstraction, respectively, based on each implementation.

4.3 Porting Methodology

The common porting methodology used across all platforms is to exploit the large amount of data-level parallelism available throughout the processing of a single IPA query. We describe the platform-specific highlights of our porting efforts in the following subsections.

4.3.1 Multicore CPU

We use the Pthread library to accelerate the kernels on the multicore platform by dividing the size of the data. Each thread is responsible for a range of data over a fixed number of iterations. This approach allows each thread to run concurrently and independently, synchronizing only at the end of the execution.

For the image matching kernels, we pre-process the input images for feature extraction by tiling the images. Each thread of the CPU is assigned one or more tiles of the input image (depending on the size of each tile). This allows us to spawn threads once at the beginning of execution and synchronize threads at the end, instead of parallelizing at a smaller granularity within the SURF algorithm, which would require multiple synchronizations between loops. However, as the tile size decreases, the number of "good" keypoints decreases, so we fix the tile size to a minimum of 50×50 per thread.

4.3.2 GPU

We use NVIDIA's CUDA library to port the Sirius components to the NVIDIA GPU. To implement each CUDA kernel, we varied and configured the GPU block and grid sizes to achieve high resource utilization, matching the input data to the best thread layout. We ported additional string manip-

ulation functions currently not supported in CUDA for the stemmer kernel.

4.3.3 Intel Phi

We port our Pthread versions to the Intel Phi platform, leveraging the ability of the target compiler to parallelize the loops on the target platform. For this, we use Intel's ICC cross-compiler. The Phi kernel is built and run directly on the target device allowing for rapid prototyping and debugging. On the Phi platform, we sweep the total amount of threads spawned in increments of 60, increasing the number of hardware threads per core. For some kernels, the maximum number of threads (with enough input data) did not always yield the highest performance. To investigate the potential of this platform to facilitate ease of programming, we use the standard programming model and custom compiler to extract performance from the platform. As such, the results represent what can be accomplished with minimal programmer effort.

4.3.4 FPGA

We use previously published details of FPGA implementations for a number of our Sirius Benchmarks in this work. However, due to limited published details for two of our workloads and to gain further insights, we design our own FPGA implementations for both GMM and Stemmer and evaluate them on a Xilinx FPGA.

GMM The major computation of the algorithm lies in three nested loops that iteratively score the feature vector against the training data. This training data comes from an acoustic model, a language model, and a dictionary in the forms of a means vector, a pre-calculated (precs) vector, a weight vector, and a factor vector. All of this data is used to generate a score for the probability of an HMM state transition. Our focus when implementing the algorithm on the FPGA was to maximize parallelization and pipeline utilization, which led to the design presented in Figure 11. This figure depicts both a core that computes the score of a single iteration of the outermost loop and a callout of a log differential unit. The log differential unit is used to fully parallelize the innermost loop, while the entire core can be instantiated multiple times to parallelize the outermost loop. Because of this, the design is highly scalable as multiple cores can be used to fill the FPGA fabric. The middle loop of the algorithm was not parallelizable, however, and is represented by the Log Summation unit. With this design, we were able to create a high throughput device with a linear pipeline.

Stemmer The Stemmer algorithm computes the root of a word by checking for multiple conditions, such as the word's suffixes or roots. Figure 12 summarizes a single step for our stemmer implementation. By taking advantage of the mutual exclusivity of test conditions, we were able to parallelize these comparisons, which allowed the FPGA to achieve a much lower latency than the original Porter algorithm. Our implementation performs multiple vector operations simultaneously to count vowels, vowel-consonant pairs, and compare suffixes. Together, these operations select the correct word shift for the specific step. We formed a single pipelined core based upon six steps dealing with the different possibilities of suffixes. We instantiate multiple cores to fill the FPGA fabric to deliver maximum performance.

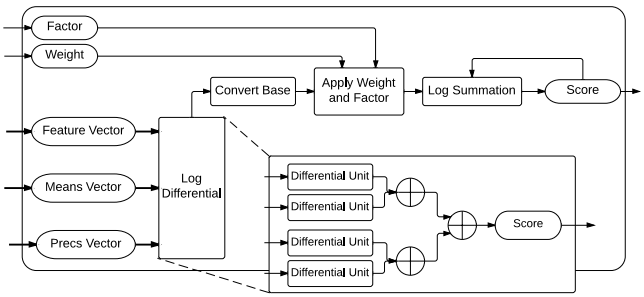


Figure 11: FPGA GMM Diagram

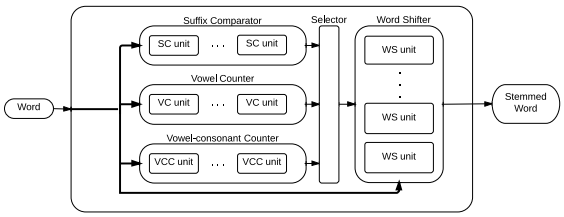


Figure 12: FPGA Stemmer Diagram

4.4 Accelerator Results

Table 5 and Figure 13 present the performance speedup achieved by the Sirius kernels running on each accelerator platform, organized by service type. For the numbers from the prior literature, we scale the FPGA speedup number to match our FPGA platform based on fabric usage and area reported in prior work. We also use numbers from the literature for kernels (Regex and CRF) that were already ported to the GPU architecture and yielded better speedups than our implementations.

4.4.1 ASR

The GMM implementation, extracted from CMU Sphinx's acoustic scoring, had the best performance on the GPU (70×) after optimizations. These custom optimizations on the GPU achieved an order of magnitude improvement by optimizing the data structure layout to ensure coalesced global memory accesses. This leveraged concurrent reads to sequential memory positions for a warp of 32 threads. In addition, it was possible to store the entire data required for the GMM in the GPU memory (2GB) during the deployment

time reducing communication between the host and device. The Phi platform did not perform as well as the GPU, indicating that the custom compiler may not have achieved the optimal data layout. The FPGA implementation using a single GMM core achieved a speedup of 56×; when fully utilizing the FPGA fabric we achieved a 169× speedup using 3 GMM cores. RWTH's DNN includes both multithreaded and GPU versions out-of-the-box. The RWTH's DNN parallelizes the entire framework (both HMM search and DNN scoring) and achieves good speedup in both cases. In the cases where we use a custom kernel or cite literature, we assume a 3.7× speedup for the HMM [35] as a reasonable lower bound.

4.4.2 QA

The NLP algorithms as a whole have very similar performance across platforms because of the nature of the workload: high input variability with many test statements causes high branch divergence. Fine tuning the stemming algorithm on the Phi to spawn 120 threads instead of the maximum and switching from allocating a range of data per thread to interlaced array accesses yields a better performance given the lower number of threads used. The FPGA stemmer implementation achieved a 6× speedup over the baseline with a single core using only 17% of the FPGA. Scaling the number of cores to fully utilize the resources of the FPGA yielded a 30× speedup over the baseline. The stemmer algorithm contains many test statements and is not well suited for SIMD operations. We attempted to improve our initial stemmer implementation for GPU by replacing most of the conditional branches with efficient XOR operations [36]. However, our fine-grained XOR-based implementation performed worse than our initial version due to additional synchronization between threads.

4.4.3 IMM

The image processing kernels achieved the best speedup on the GPU which uses heavily optimized OpenCV [16] SURF implementations yielding speedups of 10.5× and 120.5× for FE and FD, respectively. Prior work shows that FPGA yields better FE speedups but does not show similar increases for FD. The tiled multicore version yields good speedup but the performance does not scale as well on the Phi because the number of tiles is fixed, which means there is little advantage to having more threads available. The GPU version has better performance because it uses a data layout explicitly optimized for a larger number of threads.

5. Implications for Future Server Design

In this section, we investigate the performance, power and cost-efficiency trade-offs when configuring servers with different accelerator platforms for Sirius.

Table 5: Speedup of Sirius Suite Across Platforms

Service	Benchmark	CMP	GPU	Phi	FPGA
ASR	*GMM*	3.5	70.0	1.1	169.0
	DNN	6.0*	54.7*	11.2	110.5 [37]
QA	*Stemmer*	4.0	6.2	5.6	30.0
	Regex	3.9	48.0 [38]	1.1	168.2 [39]
	CRF	3.7	3.8 [40]	4.7	7.5 [41]
IMM	*FE*	5.2	10.5	2.5	34.6 [42]
	FD	5.9	120.5	12.7	75.5 [42]

* This includes DNN and HMM combined.

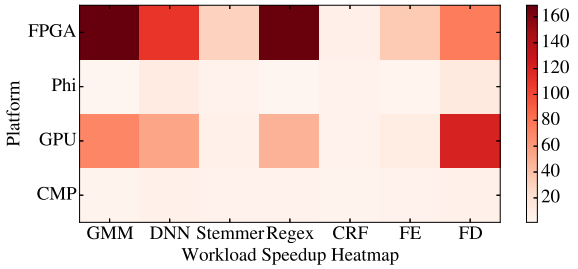


Figure 13: Heat Map of Acceleration Results

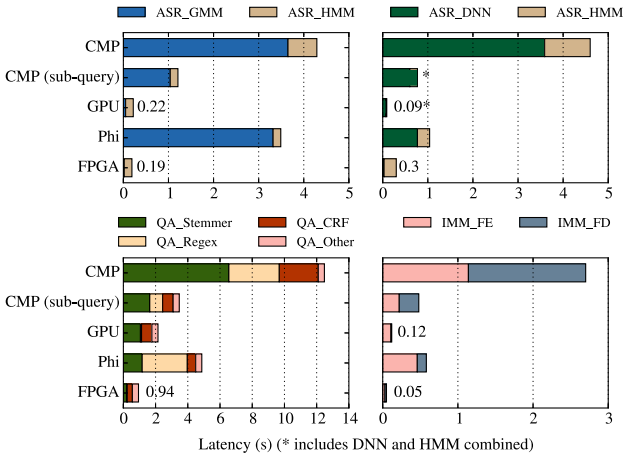


Figure 14: Latency Across Platforms for Each Service

5.1 Server Level Design

We first investigate the end-to-end latency reduction and the power efficiency achieved across server configurations for Sirius' services including ASR, QA and IMM.

5.1.1 Latency Improvement

Figure 14 presents the end-to-end query latency across Sirius' services on a single leaf node configured with each accelerator. We present both results for ASRs that use GMM/HMM and DNN/HMM as key algorithms. The latency breakdown for all hot components within a service is also presented in the figure. For QA, we focus on the NLP components comprising 88% of the cycles of QA as search has already been well studied [30].

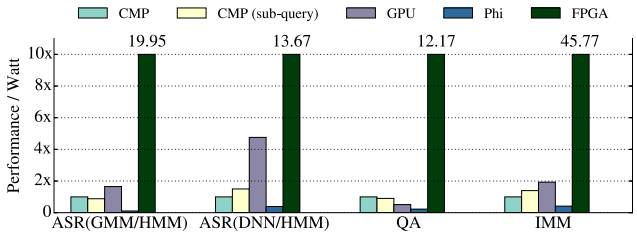


Figure 15: Performance per Watt

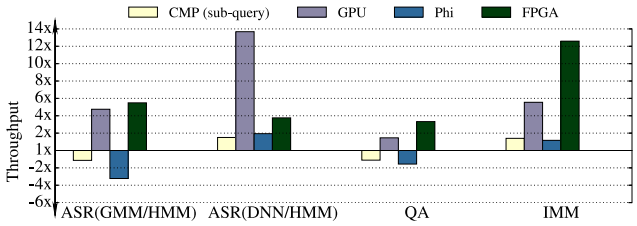


Figure 16: Throughput Across Services

Table 6: Platform Power and Cost

Platform	Power TDP (W)	Cost ($)
Intel Xeon CPU E3-1240	80	250
NVIDIA GPU GTX 770	230	399
Intel Xeon Phi 5110P	225	2,437
Xilinx Virtex-6 FPGA	22	1,795

Our baseline in this figure, CMP, is the latency of the original algorithm implementations of Sirius running on a single core of an Intel Haswell server, described in Table 3. CMP (sub-query) is our Pthreaded implementation of each service exploiting parallelism within a single query, thus reducing the single query latency. This is executed on 4 cores (8 hardware threads) of the Intel Haswell server. CMP (sub-query) in general achieves a 25% latency reduction over the baseline. Across all services, the GPU and FPGA significantly reduce the query latency. For example, the FPGA implementation of ASR (GMM/HMM) reduces the speech recognition query latency from 4.2s to only 0.19s. The FPGA outperforms the GPU for most of the services except ASR (DNN/HMM). Although Intel Phi can reduce the latency over the single core baseline (CMP), Phi is generally slower than the Pthreaded multicore baseline.

5.1.2 Energy Efficiency

Figure 15 presents the energy efficiency (performance/watt) for each accelerator platform across four services of the Sirius pipeline, normalized by the performance/watt achieved by using all cores on a multicore CPU by query-level parallelism. Here performance is defined as $1/latency$. Table 6 presents the power (TDP) for each accelerator platform. The FPGA has the best performance/watt, exceeding every other platform by a significant margin, with more than 12× energy efficiency over the baseline multicore. The GPU's performance/watt is also higher than the baseline for 3 of 4 services. Its performance/watt is worse than the baseline for QA, mainly due to its moderate performance improvement for this service.

5.2 Datacenter Design

Based on the latency and energy efficiency trade-offs for server platforms discussed in the previous section, we evaluate multiple design choices for datacenters composed of accelerated servers to improve performance (throughput) and reduce the total cost of ownership (TCO).

5.2.1 Throughput Improvement

The latency reduction shown in Figure 14 can translate to significant throughput improvement. Figure 16 presents the throughput improvement achieved using various acceleration platforms without degrading latency beyond the baseline. Similar to Figures 14 and 15, the CMP baseline executes the original Sirius workload on the Intel Haswell platform, where all four cores are utilized to serve queries, thus achieving similar throughput as CMP (sub-query level). Note that CMP's query latency is however significantly longer because CMP (sub-query level) exploits parallelism within a single query. Figure 16 demonstrates that significant latency reductions achieved by the GPU and FPGA translate to significant throughput improvement. For example, the GPU provides 13.7× throughput improvement over the baseline CMP for ASR (DNN/HMM), while the FPGA achieves 12.6× throughput for IMM. For QA, the throughput improvement across the platforms is generally more limited than other services.

Figure 17 presents the throughput improvement achieved using each acceleration platform at various load levels (the server is modeled as M/M/1 queue). Compared to Figure 16, which presents the throughput improvement at 100% load, when considering queuing effect, the lower the server load, the bigger impact latency reduction would have on throughput improvement. In other words, Figure 16 demonstrates a lower bound of throughput improvement for a queuing system. Since datacenter servers often operate at medium-to-low load, as shown in Figure 17, significant higher throughput improvement can be expected.

5.2.2 TCO Analysis

Improving throughput allows us to reduce the amount of computing resources (servers) needed to serve a given load. However, reducing the number of servers may or may not lead to reduction in the total cost of ownership of a datacenter (DC). Although reducing the machines leads to reduction on DC construction cost and power/cooling infrastructure cost, we may increase the per server capital or operational expenditure cost either by additional accelerator purchase cost or the energy cost. Here we present a cost analysis to

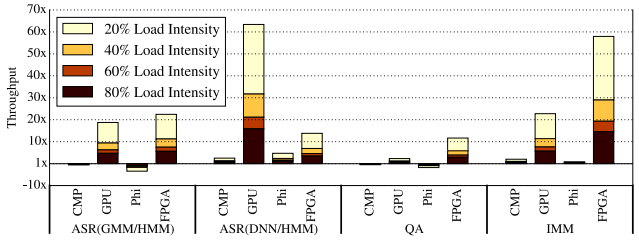


Figure 17: Throughput Improvement at Various Load Levels Modeled as M/M/1 Queue (darker is a higher load intensity for each platform)

Table 7: TCO Model Parameters [43]

Parameter	Value
DC Depreciation Time	12 years
Server Depreciation Time	3 years
Average Server Utilization	45%
Electricity Cost	$0.067/kWh
Datacenter Price	$10/W
Datacenter Opex	$0.04/W
Server Opex	5% of Capex / year
Server Price (baseline)	$2,102 [44]
Server Power (baseline)	163.6 W [44]
PUE	1.1

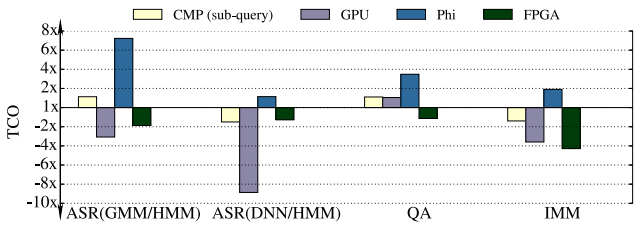


Figure 18: TCO Across Platforms for Each Service

evaluate the implication on the datacenter cost when using each accelerated server platform.

We perform our TCO analysis using the TCO model recently proposed by Google [43]. The parameters used in our TCO model are described in Table 7. The server price and power usage are based on the following server configuration based on the OpenCompute Project: 1 CPU Intel Xeon E3-1240 V3 3.4 GHz, 32 GB of RAM, and two 4TB disks [44].

Figure 18 presents the datacenter TCOs with various acceleration options, normalized by the TCO achieved by a datacenter that uses only CMPs. Overall, FPGA and GPU provide high TCO reduction. For example, GPU achieves over 8× TCO reduction for ASR(DNN) and FPGA achieves over 4× TCO reduction for IMM. We will further discuss the TCO results and use them to derive our DC designs in the next section.

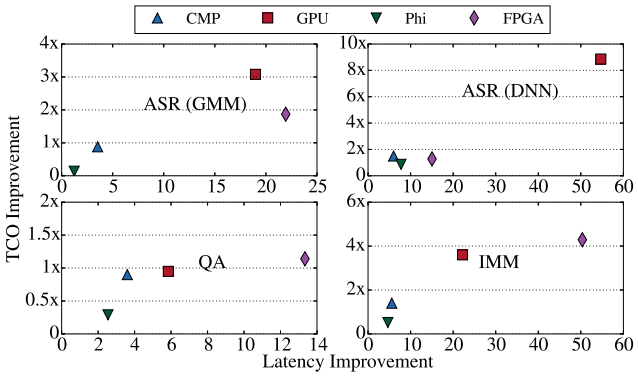


Figure 19: Trade-off Between TCO and Latency

5.2.3 Homogeneous Datacenter Design

Based on latency results from Figure 14 and TCO results from Figure 18, we first investigate the trade-offs when designing a homogeneous datacenter, that is, all servers in the datacenter have the same configuration. Homogeneous datacenters are often desirable as they minimize the management and maintenance overhead [45].

When designing a datacenter, it would be ideal to maximize performance (e.g, minimize query latency or improve throughput for a given latency constraint) and minimize the total cost of ownership. However, trade-offs may need to be made as which objective should be prioritized if both cannot be optimized by the same design. Figure 19 presents the trade-offs between the query latency improvement and the TCO improvement for each server option across four Sirius services. The x-axis presents latency improvement and the y-axis shows the TCO improvement.

As shown in the figure, FPGA achieves the lowest latency (highest latency improvement) among all accelerating platforms for 3 out of 4 services that we studied. However, the FPGA's relatively high purchase cost allows GPUs to achieve similar or higher TCO savings as FPGAs with smaller latency reduction. When the FPGA is not considered an option, the GPU achieves the optimal latency and TCO for all services. Even with the FPGA as an accelerator candidate, a GPU-accelerated datacenter provides the best latency and TCO for ASR using DNN.

Table 8 summarizes the homogeneous datacenter design for each of the main Sirius services under different conditions and optimization objectives. We present three first-order design objectives: minimizing latency, minimizing TCO with a latency constraint, and maximizing energy efficiency with a latency constraint, shown as three rows of the table. The latency constraint here is CMP (sub-query) latency shown in Figure 14. The first row (with FPGA, without FPGA, without FPGA or GPU) also shows the design constraints for the accelerator candidates.

Key Observation - *In conclusion, FPGAs and GPUs are the top 2 candidates for homogeneous accelerated datacenter designs across all three design objectives. An FPGA-accelerated datacenter allows DCs to minimize latency and*

Table 8: Homogeneous DC

	With FPGA				Without FPGA				Without {FPGA, GPU}			
	ASR (GMM)	ASR (DNN)	QA	IMM	ASR (GMM)	ASR (DNN)	QA	IMM	ASR (GMM)	ASR (DNN)	QA	IMM
Hmg-latency	FPGA				GPU				CMP			
Hmg-TCO (w/ L constraint)	GPU											
Hmg-power eff. (w/L constraint)	FPGA											

Table 9: Heterogeneous DC

	With FPGA				Without FPGA				Without {FPGA, GPU}			
	ASR (GMM)	ASR (DNN)	QA	IMM	ASR (GMM)	ASR (DNN)	QA	IMM	ASR (GMM)	ASR (DNN)	QA	IMM
Hetero-latency	FPGA	GPU (3.6x)	FPGA		GPU				CMP			
Hetero-TCO (w/ L constraint)	GPU		FPGA (20%)	FPGA (19%)								
Hetero-power eff. (w/L constraint)	FPGA											

maximize energy efficiency for most of the services and is the best homogeneous design option for those objectives. Its power efficiency is desirable for datacenters with power constraints, especially for augmenting existing filled datacenters that are equipped with capped power infrastructure support. It also improves TCO for all four services. On the other hand, FPGA-accelerated datacenters incur higher engineering cost than the rest of the platforms. For DCs where engineering cost needs to be under a certain constraint, GPU-accelerated homogeneous datacenters achieve relatively low latency and high throughput. They also achieve similar or higher TCO reduction than FPGA due to its low purchase cost. GPUs could be a desirable option over FPGAs when the high engineering overhead of FGPA implementation is a concern, especially given the quick workload churn (e.g., binaries are updated on the monthly basis) in modern datacenters.

5.2.4 Heterogeneous (Partitioned) Datacenter Design

Next, we explore the design options for partitioned heterogeneous datacenters. Because each service can run on its most suitable platform in a partitioned heterogeneous datacenter, this strategy may provide additional opportunities for further latency reduction or TCO reduction. Table 9 shows various DC design choices for different design objectives (rows), accelerator candidate sets (with FPGA, without FPGA, and without FPGA and GPU) and services (columns). The numbers in parenthesis show the improvement on the metric of the specific design objective of that row when the DC design switches from a homogeneous baseline to a heterogeneous partitioned design.

As shown in the first row of the table, when designing a partitioned heterogeneous DC for ASR, QA and IMM services, if all accelerators are considered viable candidates, GPUs can be used to optimize the latency for ASR (DNN) and achieves 3.6× latency reduction for that service compared to the homogeneous DC using FPGA across all services. Similarly, using FPGAs for QA and IMM achieves 20% and 19% TCO improvement, respectively.

Key Observation - *In conclusion, the partitioned heterogeneity in our study does not provide much benefit over the homogeneous design. The amount of benefit is certainly dependant on the workload partition across services. However, overall, most of the algorithms and services in Sirius workload exhibit a similar trend in terms of preferences for accelerators for FPGA and GPU, etc. There is also additional cost associated with managing a heterogeneous/partitioned datacenter that needs to be justifiable by the performance gain.*

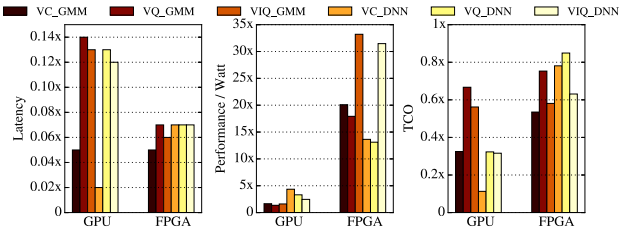


Figure 20: Latency, Energy Efficiency and TCO of GPU and FPGA Datacenters

5.2.5 Query-level Results for DC designs

In previous sections, we focused on latency, energy-efficiency and TCO trade-offs for various acceleration options across three services in Sirius. In this section, we focus on these trade-offs across three query types supported by Sirius, namely, VC, VQ and VIQ. Figure 20 presents the query latency of three query types achieved by the best two homogeneous datacenters, composed of GPU- and FPGA-accelerated servers, respectively. In addition to query latency, energy efficiency of the servers and the TCO of the datacenters to support these query types are also presented. GPU-accelerated homogeneous datacenters achieve on average 10× latency reduction, and FPGA-accelerated datacenters achieve a 16× reduction. The accelerated datacenters also reduce the TCO on average by 2.6× and 1.4×, respectively.

Figure 21 further presents the latency reduction of these two accelerated datacenters and how homogeneous accelerated datacenters can significantly reduce the scalability gap for datacenters, from the current 165× resource scaling, shown in Figure 7, down to 16× and 10× for GPU- and FPGA-accelerated datacenters, respectively.

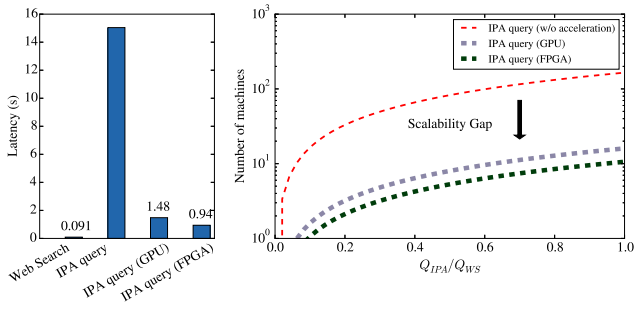


Figure 21: Bridging the Scalability Gap

6. Related Work

In addition to prior work focusing on datacenter efficiency [45–54], a heterogeneous server design [55] was proposed for speech and image recognition, where the GMM scoring and image matching algorithms were ported to hardware accelerators. However their work does not address the acceleration of NLP algorithms or DNN-based speech recognition. Custom accelerators for specific cloud applications have also been proposed, for example for memcached [56] and database systems [57] showing the growing need for specialized hardware in server applications. The Catapult project [58] at Microsoft Research has ported key components of Bing's page ranking to FPGAs. In this work, we focus on accelerating the components that make up an intelligent personal assistant focusing on their impact in the end-to-end system.

Prior work has also investigated acceleration of individual components of Sirius on various platforms. For speech recognition systems using GMM/HMM, prior work characterizes and accelerates the workload in hardware [59, 60]. In the past, GPUs have been successful in accelerating speech recognition's GMM [61] and more recently ASR was ported using a hybrid CPU-GPU approach [62]. The Carnegie Mellon *In Silicon Vox* [63] project has implemented an FPGA based GMM/HMM speech recognizer with a relatively small vocabulary. Image processing algorithms have been shown to map well to accelerators [64–66]. Key natural language processing techniques also show promising results when ported to hardware [67, 68]. Additionally, low-power accelerators for deep neural networks [69, 70] have garnered the interest of researchers as DNNs can be parallelized easily but have better accuracy compared to conventional machine learning techniques [71].

7. Conclusion

This work introduces **Sirius**, an open end-to-end intelligent personal assistant application, modeled after popular IPA services such as Apple's Siri. Sirius leverages well-established open infrastructures for speech recognition, computer vision, and question-answering systems. We use Sirius to investigate the performance, power, and cost implications of hardware accelerator-based server architectures for future data center designs. We show that GPU- and FPGA-accelerated servers can improve the query latency on average by 10× and 16×. Leveraging the latency reduction, GPU- and FPGA-accelerated servers can reduce the TCO by 2.6× and 1.4×, respectively.

8. Acknowledgment

We thank our anonymous reviewers for their feedback and suggestions. This work was partially sponsored by Google, ARM, the Defense Advanced Research Projects Agency (DARPA) under agreement HR0011-13-2-000, and by the National Science Foundation (NSF) under grants CCF-SHF-1302682 and CNS-CSR-1321047.

References

[1] Apple's Siri. https://www.apple.com/ios/siri/.

[2] Google's Google Now. http://www.google.com/landing/now/.

[3] Microsoft's Cortana. http://www.windowsphone.com/en-us/features-8-1.

[4] Smartphone OS Market Share, Q1 2014. http://www.idc.com/prodserv/smartphone-os-market-share.jsp.

[5] Google's Android Wear. www.android.com/wear/.

[6] Google's Google Glass. www.google.com/glass.

[7] ABI Research. Wearable Computing Devices, Like Apple iWatch, Will Exceed 485 Million Annual Shipments by 2018. 2013. https://www.abiresearch.com/press/wearable-computing-devices-like-apples-iwatch-will.

[8] Marti A. Hearst. 'Natural' Search User Interfaces. *Commun. ACM*, 54(11):60–67, November 2011.

[9] M. G. Siegler. Apple's Massive New Data Center Set To Host Nuance Tech. http://techcrunch.com/2011/05/09/apple-nuance-data-center-deal/.

[10] David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W Black, Mosur Ravishankar, and Alex I Rudnicky. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I. IEEE, 2006.

[11] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December 2011.

[12] David Rybach, Stefan Hahn, Patrick Lehnen, David Nolden, Martin Sundermeyer, Zoltan Tüske, Siemon Wiesler, Ralf Schlüter, and Hermann Ney. RASR - the RWTH Aachen University Open Source Speech Recognition Toolkit, 2011.

[13] Frank Seide, Gang Li, and Dong Yu. Conversational speech transcription using context-dependent deep neural networks. In *Interspeech*, pages 437–440, 2011.

[14] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. Building Watson: An Overview of the DeepQA Project

— Ferrucci — AI Magazine. *AI MAGAZINE*, 31(3):59–79, September 2010.

[15] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006*, pages 404–417. Springer, 2006.

[16] G. Bradski. *Dr. Dobb's Journal of Software Tools*, 2000.

[17] Sirius: An Open End-to-End Voice and Vision Personal Assistant. http://sirius.clarity-lab.org.

[18] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *Signal Processing Magazine*, 2012.

[19] Jeffrey Dean, Greg S. Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc V. Le, Mark Z. Mao, Marc Aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, and Andrew Y. Ng. Large scale distributed deep networks. In *NIPS*, 2012.

[20] Oscar Tckstrm, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12, 2013.

[21] Kooaba, inc. http://www.vision.ee.ethz.ch/~surf/download.html.

[22] Qualcomm Acquires Kooaba Visual Recognition Company. http://mobilemarketingmagazine.com/qualcomm-acquires-kooaba-visual-recognition-company/.

[23] G David Forney Jr. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

[24] George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(1):30–42, 2012.

[25] Xuedong Huang, James Baker, and Raj Reddy. A historical perspective of speech recognition. *Commun. ACM*, 57(1):94–103, January 2014.

[26] Vijay R. Chandrasekhar, David M. Chen, Sam S. Tsai, Ngai-Man Cheung, Huizhong Chen, Gabriel Takacs, Yuriy Reznik, Ramakrishna Vedantham, Radek Grzeszczuk, Jeff Bach, and Bernd Girod. The stanford mobile visual search data set. In *Proceedings of the Second Annual ACM Conference on Multimedia Systems*, MMSys '11, pages 117–122, New York, NY, USA, 2011. ACM.

[27] Martin F Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980.

[28] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[29] Apache nutch. http://nutch.apache.org.

[30] Michael Ferdman, Almutaz Adileh, Onur Kocberber, Stavros Volos, Mohammad Alisafaee, Djordje Jevdjic, Cansu Kaynak, Adrian Daniel Popescu, Anastasia Ailamaki, and Babak Falsafi. Clearing the clouds: A study of emerging scale-out workloads on modern hardware. In *Proceedings of the Seventeenth International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS XVII, pages 37–48, New York, NY, USA, 2012. ACM.

[31] Intel vtune. https://software.intel.com/en-us/intel-vtune-amplifier-xe.

[32] SLRE: Super Light Regular Expression Library. http://cesanta.com/.

[33] Naoaki Okazaki. CRFsuite: a fast implementation of Conditional Random Fields (CRFs), 2007. http://www.chokkan.org/software/crfsuite/.

[34] Erik F. Tjong Kim Sang and Sabine Buchholz. Introduction to the conll-2000 shared task: Chunking. In *Proceedings of the 2Nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7*, ConLL '00, pages 127–132, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[35] Jike Chong, Ekaterina Gonina, and Kurt Keutzer. Efficient automatic speech recognition on the gpu. *Chapter in GPU Computing Gems Emerald Edition, Morgan Kaufmann*, 1, 2011.

[36] A Singh, N. Kumar, S. Gera, and A Mittal. Achieving magnitude order improvement in porter stemmer algorithm over multi-core architecture. In *Informatics and Systems (INFOS), 2010 The 7th International Conference on*, pages 1–8, March 2010.

[37] Clément Farabet, Yann LeCun, Koray Kavukcuoglu, Eugenio Culurciello, Berin Martini, Polina Akselrod, and Selcuk Talay. Large-scale FPGA-based convolutional networks. *Machine Learning on Very Large Data Sets*, 2011.

[38] Giorgos Vasiliadis, Michalis Polychronakis, Spiros Antonatos, Evangelos P. Markatos, and Sotiris Ioannidis. Regular expression matching on graphics hardware for intrusion detection. In *Proceedings of the 12th International Symposium on Recent Advances in Intrusion Detection*, RAID '09, pages 265–283, Berlin, Heidelberg, 2009. Springer-Verlag.

[39] Yi-Hua E. Yang, Weirong Jiang, and Viktor K. Prasanna. Compact Architecture for High-throughput Regular Expression Matching on FPGA. In *Proceedings of the 4th ACM/IEEE Symposium on Architectures for Networking and Communications Systems*, ANCS '08, pages 30–39, New York, NY, USA, 2008. ACM.

[40] Nico Piatkowski. Linear-Chain CRF@GPU, 2011. http://sfb876.tu-dortmund.de/crfgpu/linear_crf_cuda.html.

[41] Sriram Swaminathan, Russell Tessier, Dennis Goeckel, and Wayne Burleson. A dynamically reconfigurable adaptive viterbi decoder. In *Proceedings of the 2002 ACM/SIGDA Tenth International Symposium on Field-programmable Gate Arrays*, FPGA '02, pages 227–236, New York, NY, USA, 2002. ACM.

[42] Dimitris Bouris, Antonis Nikitakis, and Ioannis Papaefstathiou. Fast and Efficient FPGA-Based Feature Detection Employing the SURF Algorithm. In *Proceedings of the 2010 18th IEEE Annual International Symposium on Field-Programmable Custom Computing Machines*, FCCM '10, pages 3–10, Washington, DC, USA, 2010. IEEE Computer Society.

[43] Luiz Andre Barroso, Jimmy Clidaras, and Urs Holzle. *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, Second Edition*. Synthesis Lectures on Computer Architecture. 2013.

[44] Thinkmate high performance computing, 2014. http://www.thinkmate.com/system/rax-xf2-1130v3-sh.

[45] Jason Mars and Lingjia Tang. Whare-map: Heterogeneity in homogeneous warehouse-scale computers. In *ISCA '13: Proceedings of the 40th annual International Symposium on Computer Architecture*. IEEE/ACM, 2013.

[46] Jason Mars, Lingjia Tang, Robert Hundt, Kevin Skadron, and Mary Lou Soffa. Bubble-up: Increasing utilization in modern warehouse scale computers via sensible co-locations. In *Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, MICRO-44, pages 248–259, New York, NY, USA, 2011. ACM.

[47] Jason Mars, Lingjia Tang, Kevin Skadron, Mary Lou Soffa, and Robert Hundt. Increasing utilization in modern warehouse-scale computers using bubble-up. *IEEE Micro*, 32(3):88–99, May 2012.

[48] Lingjia Tang, Jason Mars, Xiao Zhang, Robert Hagmann, Robert Hundt, and Eric Tune. Optimizing google's warehouse scale computers: The numa experience. In *Proceedings of the 2013 IEEE 19th International Symposium on High Performance Computer Architecture (HPCA)*, HPCA '13, pages 188–197, Washington, DC, USA, 2013. IEEE Computer Society.

[49] Lingjia Tang, Jason Mars, Wei Wang, Tanima Dey, and Mary Lou Soffa. Reqos: Reactive static/dynamic compilation for qos in warehouse scale computers. In *Proceedings of the Eighteenth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, ASPLOS '13, pages 89–100, New York, NY, USA, 2013. ACM.

[50] Hailong Yang, Alex Breslow, Jason Mars, and Lingjia Tang. Bubble-flux: Precise online qos management for increased utilization in warehouse scale computers. In *Proceedings of the 40th Annual International Symposium on Computer Architecture (ISCA)*, ISCA '13, pages 607–618, New York, NY, USA, 2013. ACM.

[51] Yunqi Zhang, Michael Laurenzano, Jason Mars, and Lingjia Tang. Smite: Precise qos prediction on real system smt processors to improve utilization in warehouse scale computers. In *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, MICRO-47, New York, NY, USA, 2014. ACM.

[52] Michael Laurenzano, Yunqi Zhang, Lingjia Tang, and Jason Mars. Protean code: Achieving near-free online code transformations for warehouse scale computers. In *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, MICRO-47, New York, NY, USA, 2014. ACM.

[53] Vinicius Petrucci, Michael A. Laurenzano, Yunqi Zhang, John Doherty, Daniel Mosse, Jason Mars, and Lingjia Tang. Octopus-man: Qos-driven task management for heterogeneous multicore in warehouse scale computers. In *Proceedings of the 2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*, HPCA '15, Washington, DC, USA, 2015. IEEE Computer Society.

[54] Chang-Hong Hsu, Yunqi Zhang, Michael A. Laurenzano, David Meisner, Thomas Wenisch, Lingjia Tang, Jason Mars, and Ron Dreslinski. Adrenaline: Pinpointing and reigning in tail queries with quick voltage boosting. In *Proceedings of the 2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*, HPCA '15, Washington, DC, USA, 2015. IEEE Computer Society.

[55] Ravi Iyer, Sadagopan Srinivasan, Omesh Tickoo, Zhen Fang, Ramesh Illikkal, Steven Zhang, Vineet Chadha, Paul M. Stillwell Jr., and Seung Eun Lee. Cogniserve: Heterogeneous server architecture for large-scale recognition. *IEEE Micro*, 31(3):20–31, 2011.

[56] Kevin Lim, David Meisner, Ali G. Saidi, Parthasarathy Ranganathan, and Thomas F. Wenisch. Thin servers with smart pipes: Designing soc accelerators for memcached. In *Proceedings of the 40th Annual International Symposium on Computer Architecture*, ISCA '13, pages 36–47, New York, NY, USA, 2013. ACM.

[57] Onur Kocberber, Boris Grot, Javier Picorel, Babak Falsafi, Kevin Lim, and Parthasarathy Ranganathan. Meet the walkers: Accelerating index traversals for in-memory databases. In *Proceedings of the 46th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO-46, pages 468–479, New York, NY, USA, 2013. ACM.

[58] Andrew Putnam, Adrian Caulfield, Eric Chung, Derek Chiou, Kypros Constantinides, John Demme, Hadi Esmaeilzadeh, Jeremy Fowers, Gopi Prashanth Gopal, Jan Gray, Michael Haselman, Scott Hauck, Stephen Heil, Amir Hormati, Joo-Young Kim, Sitaram Lanka, Jim Larus, Eric Peterson, Simon Pope, Aaron Smith, Jason Thong, Phillip Yi Xiao, and Doug Burger. A reconfigurable fabric for accelerating large-scale datacenter services. In *41st Annual International Symposium on Computer Architecture (ISCA)*, June 2014.

[59] Rajeev Krishna, Scott Mahlke, and Todd Austin. Architectural optimizations for low-power, real-time speech recognition. In *Proceedings of the 2003 International Conference on Compilers, Architecture and Synthesis for Embedded Systems*, CASES '03, pages 220–231, New York, NY, USA, 2003. ACM.

[60] Binu Mathew, Al Davis, and Zhen Fang. A low-power accelerator for the sphinx 3 speech recognition system. In *Proceedings of the 2003 International Conference on Compilers, Architecture and Synthesis for Embedded Systems*, CASES '03, pages 210–219, New York, NY, USA, 2003. ACM.

[61] Paul R. Dixon, Tasuku Oonishi, and Sadaoki Furui. Harnessing graphics processors for the fast computation of acoustic likelihoods in speech recognition. *Comput. Speech Lang.*, 23(4):510–526, October 2009.

[62] Jungsuk Kim, Jike Chong, and Ian R. Lane. Efficient On-The-Fly Hypothesis Rescoring in a Hybrid GPU/CPU-based Large Vocabulary Continuous Speech Recognition Engine. In *INTERSPEECH*. ISCA, 2012.

[63] Edward C. Lin, Kai Yu, Rob A. Rutenbar, and Tsuhan Chen. A 1000-word Vocabulary, Speaker-independent, Continuous Live-mode Speech Recognizer Implemented in a Single FPGA. In *Proceedings of the 2007 ACM/SIGDA 15th International Symposium on Field Programmable Gate Arrays*, FPGA '07, pages 60–68, New York, NY, USA, 2007. ACM.

[64] J Hauswald, T Manville, Q Zheng, R Dreslinski, C Chakrabarti, and T Mudge. A hybrid approach to offloading mobile image classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 8375–8379. IEEE, 2014.

[65] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: an efficient alternative to sift or surf. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2564–2571. IEEE, 2011.

[66] Tung H Dinh, Dao Q Vu, Vu-Duc Ngo, Nam Pham Ngoc, and Vu T Truong. High throughput fpga architecture for corner detection in traffic images. In *Communications and Electronics (ICCE), 2014 IEEE Fifth International Conference on*, pages 297–302. IEEE, 2014.

[67] Yuliang Sun, Zilong Wang, Sitao Huang, Lanjun Wang, Yu Wang, Rong Luo, and Huazhong Yang. Accelerating frequent item counting with fpga. In *Proceedings of the 2014 ACM/SIGDA International Symposium on Field-programmable Gate Arrays*, FPGA '14, pages 109–112, New York, NY, USA, 2014. ACM.

[68] Jan Van Lunteren, Christoph Hagleitner, Timothy Heil, Giora Biran, Uzi Shvadron, and Kubilay Atasu. Designing a programmable wire-speed regular-expression matching accelerator. In *Proceedings of the 2012 45th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO-45, pages 461–472, Washington, DC, USA, 2012. IEEE Computer Society.

[69] Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, and Olivier Temam. DianNao: A Small-footprint High-throughput Accelerator for Ubiquitous Machine-learning. In *Proceedings of the 19th International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '14, pages 269–284, New York, NY, USA, 2014. ACM.

[70] Hadi Esmaeilzadeh, Adrian Sampson, Luis Ceze, and Doug Burger. Neural acceleration for general-purpose approximate programs. In *Proceedings of the 2012 45th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO-45, pages 449–460, Washington, DC, USA, 2012. IEEE Computer Society.

[71] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6645–6649. IEEE, 2013.