TODO LIST

Shahbaz: add missing citations once the paper is ready. 1
Nick: to work on para below.... incorp with privilege modes. 12

The nanoPU: From Microservices to Nanoservices

Paper #52

12 Pages Body, 14 Pages Total

ABSTRACT

Microservices have been shown to speedup a broad class of applications by employing fine-grained parallelism using serverless tasks that run for just a few milliseconds. This paper takes this approach to the extreme: Our goal is to enable much finer-grained parallelism, that we call *nanoservices*, lasting less than 1μs. This is primarily a networking problem, not a computer architecture problem. Our nanoservice Processing Unit (nanoPU), runs on a slightly-modified CPU with a very low latency network interface (NIC). The nanoPU is a domain-specific processor optimized to run sub-microsecond, cache-resident RPC "nanotasks".

The nanoPU, as a new networking-optimized compute platform, helps make nanoservices a reality. It contains a fast path from the network directly to a CPU core, minimizing communication latency, and it delegates thread scheduling along with transport logic to the NIC, minimizing tail RPC-completion times and per-message overheads.

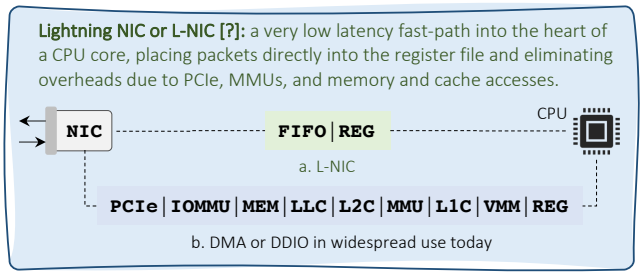
We built a prototype nanoPU on top of an open-source RISC-V CPU and evaluated its performance for a suite of real nanoservice applications using cycle-accurate hardware simulations.

The nanoPU reduces the average latency for an RPC request-response from the best reported (hundreds of microseconds) to below 1μs for a specific class of applications. We demonstrate NFV-like streaming applications running 3.5-4× faster with nanoPU

We believe that this class of applications will become increasingly popular, as many distributed applications have inherent parallelism that cannot be harnessed today because of large, unpredictable RPC-completion times and substantial per-message overheads.

1 INTRODUCTION

Cloud data centers allow thousands of client applications to share millions of CPU cores. Many of these client applications harness CPU cores using *coarse*-grained parallelism, such as MapReduce [21], Hadoop [6] and client-server web applications. Other, more specialized applications use *fine*-grained parallelism to employ hundreds of thousands of CPUs (e.g., search [12], HPC [61]). It is well-known to be difficult to write applications using fine-grained parallelism, and so the technique is generally reserved as a last resort, for when applications must respond quickly (e.g., search or ads), or must complete in a reasonable time (e.g., a large HPC physical simulation [2, 27, 33] or overnight ML training [1]). Recently,



as applications have grown ever larger (e.g., mushrooming ML training data, big data analytics, and social media), the speed of individual cores has grown ever more slowly [29]. Increased parallelism seems inevitable.

Some recent products and open-source projects have used fine-grain modules, deployed as microservices (e.g., AWS Lambda [8], Azure Serverless Computing [10], Google Cloud Functions [26], and OpenFaaS [52]) to demonstrate impressive speedup of parallel applications, including video encoding [23], video compression [5], and face recognition [16], by successfully employing thousands of CPU cores in parallel. But, as we will see, this level of parallelism still yields far from the fastest possible execution time.

In the limit, if we segment an application into its smallest, indivisible software modules, *and if communication is instantaneous*, then an application can complete in the execution time of the longest sequence of serially-dependent modules. But, communication is not instantaneous—far from it. Many researchers have reported the (surprisingly) high communication latency between two cores, e.g., a request-response time of 0.7 ms [39, 55]. Intuitively, if it takes 0.7 ms to ask another CPU to execute a function, then it is only worth doing if we have more than 0.7 ms of work to do; otherwise we might as well execute the code locally.

For example, consider a physical simulation of one million objects, where each object needs to update its ten neighbors each time it moves. If each object executes 10 μs of code to calculate its move (about 40,000 instructions on a modern CPU), but takes 1 ms to tell its neighbors, then latency dominates the computation, and it makes sense for 10–100 objects to share one core. If instead, an RPC call could complete in less than 1 μs, then it makes sense for each object to run on its own core, and the simulation can complete one hundred times sooner.

More generally, consider a distributed application consisting of N potentially parallelizable tasks, each of which can

Shahbaz: add missing citations once the paper is ready.

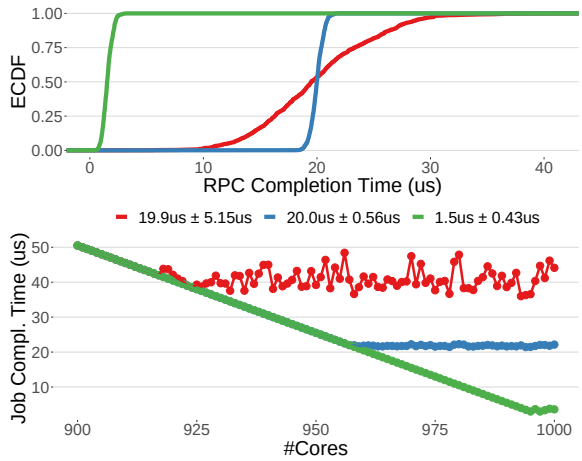


Figure 1: Total job completion time of an example application with 1000 independent tasks (each with 500 μs processing time) for varying number of cores (i.e., one per RPC) under different RPC completion time distributions.

be executed locally (in x μs) or on a remote core via RPC (in $r \cdot x$ μs, where $r \geq 1$). If the degree of parallelism, K, is the number of issued RPCs (i.e., the number of remote cores), we minimize job-completion time when $(N - K) \cdot x = r \cdot x$ (i.e., $K = N - r$). Further increasing the degree of parallelism, K, yields no benefit; if we send more RPCs, the local core just sits idle waiting for the RPCs to complete. For example, if $r = 10$ (i.e., RPCs take 10 times as long as local execution) then we should send $N - 10$ RPCs while executing 10 tasks locally. If we reduce latency so that $r = 1$, the total job-completion time is reduced by 90%.

In practice, RPC completion times, f_i, are not fixed, but characterized by a probability distribution, $f_i \in \mathcal{F}$. The total job completion time is $(N - K) \cdot x = \max f_i$ where $1 \leq i \leq K$. Thus, if we want to improve performance, we must reduce the RPC tail latency ($\max f_i$). This observation is not new and has been the motivation for much recent work [18, 38, 54].

The goal of our research is to design, build, and evaluate a domain-specific processor, which we call a *nanoPU*, optimized for minimizing the completion time of highly-parallel, compute-intensive jobs. The nanoPU is designed for massive scale-out computing. Specifically, millions of nanoPUs could be deployed to run what we call *nanoservices*: highly distributed, compute-intensive applications that process RPC requests in under 1 μs using fine-grained, cache-resident threads.

If we are to minimize the average and tail latency of an RPC, we must minimize the latency at every step of the way. Starting from when a CPU thread issues an RPC, we must minimize latency through the network stack, through the cache, DRAM and DMA subsystem, through the NIC and onto the wire; across Ethernet links and through switches; and then from the destination NIC through the DMA, DRAM, and cache subsystem; through the network stack, and then wait for the RPC request's target thread to be scheduled for execution by the operating system and for the application to read and process the request. Most prior work has focused on one aspect of an RPC's latency–for example, low-latency transport layers [4, 28, 47] to reduce network congestion and minimize latency through switches, or thread scheduling to make sure an incoming RPC request starts executing promptly [38, 54]. Our goal is to reduce RPC response times by at least an order of magnitude from tens or hundreds of microseconds down to approximately 1 μs. This is mostly a networking problem, one that leads us to tackle three questions:

(1) **Minimizing *average* RPC completion time:** How can we minimize the time that a network message spends between the wire and an application thread? This requires minimizing the time through hardware and software. Our approach is to build upon the extremely low-latency *Lightning NIC* (L-NIC), described in [31], in which messages bypass the traditional networking stack, as well as the DMA, DRAM and cache hierarchy entirely, and are placed directly into the CPU register file. An arriving RPC can start execution in under 100 ns (not including the MAC).
(2) **Minimizing *tail* RPC completion time:** How can we limit resource contention for the CPU, memory and network, and minimize the cost of context switching between threads? These are the primary causes of variance and hence tail RPC completion times. Our approach is to design a thread scheduler into the NIC hardware, as well as to enable the NIC to implement a very low latency hardware transport layer, such as Homa [47] or NDP [28].
(3) **Making it practical:** How can such an extreme approach to hardware and software be practically deployed, with minimal disruption? We have designed and implemented a hardware prototype, based on a RISC-V CPU [7], that includes a 100 Gb/s Ethernet L-NIC CPU interface and a hardware thread scheduler, and we report performance results for real nanoservice applications.

Figure 1, while based on synthetic numbers, illustrates what we aim to achieve. The first graph shows various distributions of RPC completion times. The red graph is intended to be representative of the current state-of-the-art RPC completion times [39]. The blue graph demonstrates the impact on total application runtime when the standard deviation of the RPC completion time distribution is reduced by an order of magnitude. If we can drive down the average *and*

the standard deviation, represented by the green line, then the job-completion time falls linearly as we add cores, until eventually the per-RPC overhead limits parallelism.

In summary, we make the following contributions:

- A new compute framework, called nanoservices (§2), for highly parallelizable, compute-intensive distributed applications.
- The design and implementation of a nanoPU (§3) with a 100 Gb/s NIC datapath (§3.1), a message-based interface directly between the NIC and CPU register file (§3.2) and a NIC-driven thread scheduler (§3.3).
- An overview of the hardware mechanisms designed to support a low-latency transport protocol on the NIC (§3.4).
- An open-source nanoPU prototype (§4), based on a RISC-V Rocket Core [7].
- An evaluation of our nanoPU prototype against a traditional RISC-V CPU with a DMA-based NIC, for a suite of nanoservice applications (§5). Our evaluation demonstrates that nanoPU reduces both average and tail RPC-completion times by about 2–10× and 20×, respectively.

2 NANOSERVICES

Nanoservices are a new framework for building compute-intensive, distributed applications with the following three characteristics:

C1: Each application is divisible into compute-intensive *nanotasks*, managed by a light-weight *nanokernel*.
C2: Each nanotask is able to process and respond to network messages in under 1 μs.
C3: The working set of each nanotask fits in the L1 cache.[1]

Figure 2 highlights the key differences between nanoservices and two other major application development frameworks: monoliths and microservices. We believe that many applications can benefit from the nanoservices framework, including applications such as physical simulations [27, 56], neural-network inference and training [1], graphics processing [25], and state-space search algorithms [65]. As we will show in §5, total application runtime has the potential to be reduced by about three orders of magnitude if they are implemented using nanoservices.

The central idea is as follows. A nanotask must execute in the shortest, and most deterministic time possible. This implies that *everything* the nanotask needs must sit in registers and the L1 cache: instructions in the L1 instruction cache, and data (in-progress messages, variables and program state) in registers or the L1 data cache. Cache misses are expensive, and the goal is to avoid them—L2 cache and external memory

[1] L1 cache is typically 256 kB–1 MB at the time of writing, including the data and instruction cache.

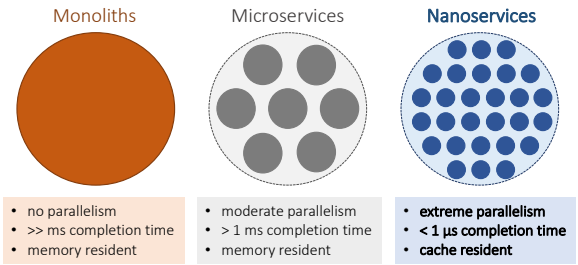


Figure 2: A comparison of key characteristics of *nanoservices* vs. monoliths and microservices.

are, respectively, ten and one hundred times slower than the L1 cache [20].

Building nanoservices requires more than simply dividing applications into nanotasks that access small amounts of data. We also need to consider how to quickly and predictably transfer message data between the wire and an application thread, and how to efficiently schedule nanotasks on the available processor cores. It is not practical to run nanoservices on today's systems, which were designed to support applications that process RPCs in milliseconds, or at best tens of microseconds [39, 55], not hundreds of nanoseconds. Addressing these problems therefore requires hardware changes, which we address with our proposed design, the nanoPU, in the following section.

3 THE NANOPU

The nanoPU is a new domain-specific processor optimized for running nanotasks—with low average and peak latency—for compute-intensive distributed applications based on nanoservices. A nanoPU consists of one or more CPU cores and one or more low-latency NICs. The CPU cores are slightly-modified cores; our design is based on the popular, open-source RISC-V ISA [59]. The low-latency NIC is inspired by the recently proposed *Lightning NIC (L-NIC)* [31], a novel approach that terminates the transport layer in hardware, then delivers message data right into the registers at the heart of the CPU core. The L-NIC approach minimizes latency (and unpredictability) by bypassing DMA, cache, and DRAM entirely. Our nanoPU design also adds a novel hardware thread scheduler, to minimize the time (and variability) from when a message arrives until the thread starts processing it.

Figure 3 shows a high-level block diagram of a nanoPU. This particular example shows one network interface shared by three cores, but in general the nanoPU is designed to work with any ratio of NICs to cores. Depending on the need, it may make sense to build nanoPUs with one core per NIC (e.g., for small embedded systems), all the way to hundreds of cores per NIC. For example, it would be practical today to design a chip with 500 cores [3, 13] and over a hundred 100 Gb/s

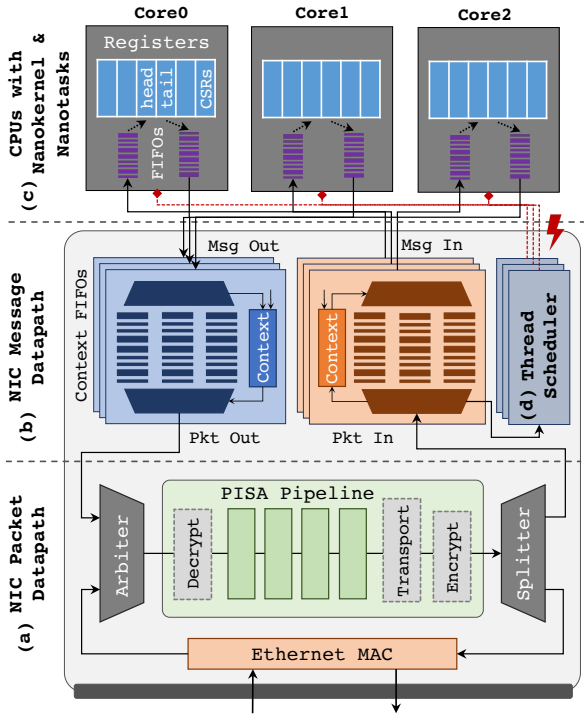


Figure 3: The nanoPU architecture. The NIC includes an event-driven PISA pipeline, and provides an RPC message abstraction to a thread via dedicated RX/TX registers in the CPU. CPU cores run a nanokernel and user nanotasks.

Ethernet interfaces;[2] in this example, the ratio would be five cores per NIC. Of course, many other ratios are possible and the ideal ratio depends on the application, technology, and economics.

The nanoPU is a domain-specific processor. With Moore's Law slowing down, new domain-specific processors are being widely used as accelerators for specific, high-volume workloads, such as graphics [51], machine-learning [1] and networking [15]. While the nanoPU is not nearly as radical a departure from a general-purpose CPU as, say, a GPU, TPU or programmable switch (after all, our design relies heavily on an existing core), the nanoPU shares the approach of tailoring the chip design for a specific class of applications.[3]

If nanoPU is so fast, why are not all CPUs designed this way? It is because general-purpose CPUs are optimized for general-purpose workloads, most of which are memory-intensive. They are "load-store" machines, with memory as a first-class citizen. Arriving and departing packet data must pass through the memory subsystem first, on its way in and out of the CPU. Our approach is, instead, to make network messages first-class citizens in their own right, independent of memory. Network messages arrive directly into the CPU, without needing to make their way through the memory hierarchy. General-purpose CPUs put memory in the network and attach compute to memory, which makes perfect sense for memory-intensive applications. Our approach in the nanoPU is to tightly couple compute directly to the network, and then attach memory on the side as needed.

The nanoPU has the following key characteristics that we visit, in turn, in the next few subsections:

- **NIC Packet Datapath:** A programmable event-driven PISA "match action unit" (MAU) pipeline [30] on the NIC to process packet headers as they arrive and depart over the network, terminate tunnels, encrypt/decrypt and compress/decompress data; and a low-latency transport protocol in hardware, such as Homa [47] or NDP [28].
- **NIC Message Datapath:** A very low latency path—just a few clock cycles—from the network right to the very heart of the CPU core—its register file. This reduces latency by 1–2 orders of magnitude; we do not believe there is a lower latency path to a running thread.
- **Hardware Thread Scheduling:** In addition to moving network messages into the CPU quickly, the nanoPU must also make sure that the appropriate application thread is running on the core so that it can start processing messages promptly. For this, nanoPU includes a very low-latency, thread scheduler on the NIC and a light-weight nanokernel on the CPU.

These characteristics together enable the nanoPU to process RPCs quickly and predictably making it the ideal platform for building compute-intensive, distributed applications.

3.1 The NIC Packet Datapath

The NIC packet datapath (Figure 3a) processes packets as they enter and leave the network. The centerpiece of the NIC packet datapath is an event-driven PISA pipeline [30]. The original PISA architecture, proposed in the RMT chip [15] and later used by Tofino [63], is designed for mostly-stateless match-action processing of packet headers; for example, for lookups, encapsulation, tunnels and telemetry. Basic PISA only supports one type of event: the arrival of new packets. *Event-driven* PISA enhances the basic model to support other event types, such as packet drops, timers and state-dependent events. Section 3.4 describes how an event-driven PISA pipeline can directly support transport protocols in

[2]Commercial switch chips exist with 128 × 100 Gb/s Ethernet MACs today.
[3]As an aside, it is only possible to consider prototyping a nanoPU in a university because of two recent trends: RISC-V provides a remarkably stable starting point, and the narrow performance gap between the leading edge process (currently 7 nm) and the closest already-paid-for process (16 nm) makes it economically feasible to build an interesting nanoPU.

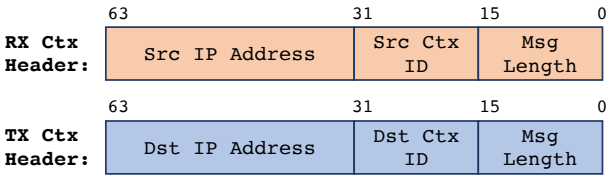


Figure 4: NIC message (context) header formats.

hardware, thus offloading per-packet processing from the CPU.

The PISA pipeline also performs protocol processing: depending on the loaded P4 program [14], it might add and remove Ethernet, IP, VXLAN, GRE, INT [41] and transport headers. We also program it to add or remove a small application-specific per-context header containing an IP address, a context ID, and the message length (Figure 4).

As others have noted, a P4-programmable PISA pipeline can also be used to accelerate some applications by offloading processing from the CPU (e.g., memory and disk caches [35], load-balancers [45], consensus protocols [34] and firewalls [58]). The L-NIC paper [31] describes how a PISA pipeline can accelerate nanoservices to search the Othello state-space.

3.2 The NIC Message Datapath

The nanoPU's NIC message datapath (Figure 3b), is responsible for assembling arriving packets into a self-contained RPC message, then queues it in a per-context FIFO until its destination thread is scheduled on a core by a nanokernel to process it. The message then flows, one word at a time, into the core's register file. In the egress direction, messages sent by applications' threads are broken into packets before being sent to the NIC packet datapath.

We make two small modifications to the CPU core:

- Two general-purpose registers (GPRs) are now reserved for a special purpose: one is the `HEAD` of the network receive queue, the other is the `TAIL` of the network transmit queue. Applications must be compiled to avoid using reserved GPRs for temporary storage. Fortunately, `gcc` makes it easy to reserve registers via command-line options [53].
- New control status registers (CSRs) are added for out-of-band communication between the CPU and the NIC. These are used to configure the NIC with context IDs and to enable NIC-driven thread scheduling.

The NIC message datapath maintains transmit and receive FIFOs for each context (i.e., thread) that is currently pinned to each core. It ensures that the currently running thread only sees (i.e., reads from and writes to) its own per-context FIFO via the `HEAD` and `TAIL` GPRs.

```
// Simple loopback & increment nanoservice
entry:
  // Register context ID & priority with NIC
  csrwi lcurcontext, 0
  csrwi lcurpriority, 0
  csrwi lniccmd, 1

// Wait for a message to arrive
wait_msg:
  csrr  a5, lmsgsrdy
  bnez  a5, loopback_plus1_16B
idle:
  csrwi lidle, 1
  csrr  a5, lmsgsrdy
  beqz  a5, idle

// Loopback and increment 16B message
loopback_plus1_16B:
  mv t6, t5 // copy ctx hdr from head to tail
  addi t6, t5, 1 // first data word + 1
  addi t6, t5, 1 // second data word + 1
  j wait_msg // wait for the next message
```

Listing 1: Loopback with increment: a simple RISC-V assembly program for the nanoPU that waits for a 16 B **message to arrive, increments each word of the message, and returns it to the sender.**

The NIC exposes a *message* interface to applications running on the core, instead of the more traditional packet interface. This makes the nanoPU NIC hardware responsible for the transport layer, which includes breaking messages into packets, ensuring reliable delivery across the network, and performing message reassembly at the receiving end. To this end, messages between the NIC and core carry a small 8-byte header whose format is shown in Figure 4.

A nanoPU program (loopback-increment). To illustrate how software on the nanoPU interacts with the NIC, Listing 1 shows a simple loopback-increment program in RISC-V assembly language. The program continuously reads 16-byte messages (consisting of two 8-byte integers) from the network, increments the integers, and sends the messages back to their sender. The program is described in more detail below.

The `entry` procedure registers a context ID and its priority with the NIC by first writing a value to both the `lcurcontext` and `lcurpriority` CSRs, then subsequently writes the value 1 to the `lniccmd` CSR. The `lniccmd` CSR is a bit vector used by software to send commands to the NIC—in this case, it is used to tell the NIC to allocate an RX and a TX queue for context ID = 0 with priority level = 0. The `lniccmd` CSR can

also be used to remove context IDs or to update the priority level.[4]

The `wait_msg` procedure waits for a message to arrive in the RX queue by polling the `lmsgsrdy` CSR until it is set by the NIC, indicating that the context has messages to process. While it is waiting, the application tells the NIC that it is idle by setting the `lidle` CSR during the polling loop. The NIC thread scheduler uses the idle signal to evict waiting threads in order to schedule a new thread that has messages waiting to be processed.

The `loopback_plus1_16B` procedure simply swaps the source and destination addresses by moving the context header (the first word of every message) from the head register (`t5`) to the tail register (`t6`), shown on line 20 (Listing 1). It then increments every integer in the received message, appends them to the message being transmitted, and waits for the next message to arrive.

Applications that use variable-length messages can use the message length (in the context header) to read the correct number of words from the network RX queue. If an application reads an empty RX queue, the resulting behavior is undefined—similar to reading an uninitialized variable.

3.3 NIC-Driven Thread Scheduling

In some specialized applications, thread scheduling might not be needed. Big HPC applications, for example, might pin every nanotask to a dedicated core for the duration of a run. But this would not work for a cloud provider who needs more economical sharing of nanoPUs by multiple nanoservice applications. Nanotask threads will therefore need to be switched in and out frequently, and context switches must be extremely fast. Otherwise we will lose all the low latency benefits of nanoservices.

If nanoPU relied on software scheduling it would be too slow. The fastest best-of-breed software schedulers use 5μs timer interrupts [38, 54] which are much too slow for our $< 1\mu$s nanotasks. And so, instead, the NIC hardware (Figure 3d) decides which thread to run next, and informs the nanoPU's minimal operating system—the nanokernel. The NIC keeps track of the highest-priority nanotask with messages to process and interrupts the CPU to initiate a context switch under two conditions:

(1) A complete message arrives for a higher priority context than the one currently running on the core.
(2) The current context tells the NIC it is idle, and the NIC has messages for another context to process. This condition prevents an idle high-priority context from starving a low-priority context with messages waiting to be processed.

[4]Registering a context ID with the NIC is roughly equivalent to opening a socket on a modern OS.

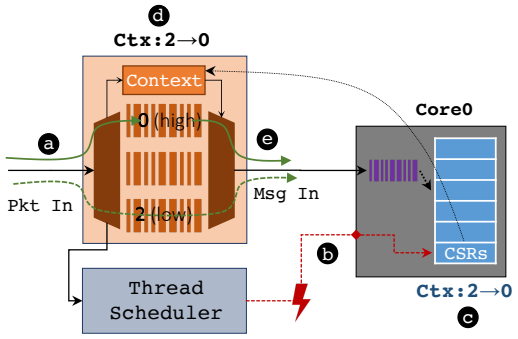


Figure 5: The steps that occur when a low-priority thread running on a core is preempted by a higher-priority thread.

Figure 5 shows what happens when a message arrives for a higher-priority context than the one currently running on the core. In step ⓐ, the arriving message is enqueued into the RX queue for the high-priority thread. In step ⓑ and ⓒ, the thread scheduler tells the CPU that a high-priority message is waiting by updating the `ltargetcontext` CSR ($2 \rightarrow 0$) and firing an interrupt. The interrupt causes a trap into the nanokernel, which then reads the `ltargetcontext` CSR and performs a switch to the high-priority context in step ⓓ, updating the `lcurcontext` CSR. At this point, the high-priority context is running on the core and is able to process the message in its RX queue ⓔ. When the high-priority context finishes processing the message, it tells the NIC that it is now idle by writing to a dedicated CSR called `lidle`. The NIC then initiates a context switch back to the low-priority context because it still has messages to process.

3.4 Terminating Transport in the NIC

If the nanoPU is to meet our aggressive latency target of $< 1\mu$s from when an RPC message arrives over Ethernet until a response is sent back out, there is clearly no time to run a traditional transport networking stack in software. Therefore, nanoPU runs a low-latency transport protocol in hardware, in the NIC, with support from the NIC's programmable pipeline. Note that transport protocols for low-latency applications (e.g., DCQCN for RDMA) are already fully implemented in hardware [32, 44] on state-of-the-art high-speed NICs available in today's market, vouching for the feasibility of a fully-offloaded transport protocol running in hardware.

We, however, don't mandate a specific transport protocol in this paper because we seem to be entering an era when different cloud providers prefer different transport protocols [42, 46, 66]. Instead we aim to provide some choices, within our tight latency budget.

The NIC therefore places minimal constraints on the transport protocol, while providing programmable hardware blocks

to allow some choice by the network owner. We assume that the transport layer provides a reliable message abstraction to each thread, as well as network congestion control. It therefore must handle retransmissions and decide when packets should be sent. To guide our design, we assume it must be possible for the NIC to be programmed to implement Homa [47] and NDP [28]. Between them, they mandate most of the building blocks we need, including reliable transmission, immediate startup rate, receiver driven scheduling, a message abstraction to the CPU, and data-trimming (in the switches).

We observe that realizing a transport protocol in hardware requires the following functions in the programmable pipeline of a NIC.

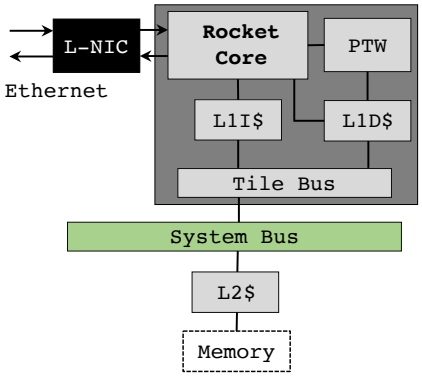
- Timers and timer-based event processing logic to realize various timer-based state transitions, such as retransmissions and timeouts.
- Pacers to rate-limit individual flows.
- State machines to maintain per-flow state, including current rate or congestion-window size, sequence and acknowledgment numbers, connection status, counters, etc.
- Packetization/retransmission buffer to break a message into packets and hold packets until they are acknowledged by a receiver.
- Reordering buffer to handle out-of-order packets.
- Packet generators to realize receiver-driven transport protocols that keep generating credit packets for senders.

The event-driven PISA pipeline [30] already provides most of the mechanisms we need for sophisticated stateful operations (e.g., state machines, timer events and packet generation), and can be extended to add support for message reassembly and retransmit buffers [57, 60]. Because nanoservice messages sizes will be very small, the amount of SRAM needed to realize these buffers will be sufficiently small for a cost- and power-efficient hardware implementation. With these mechanisms, we believe the nanoPU can run a complete transport stack on the NIC with very low latency. While we have a design for each of these building blocks, we will explain the details of those in a follow-up paper.

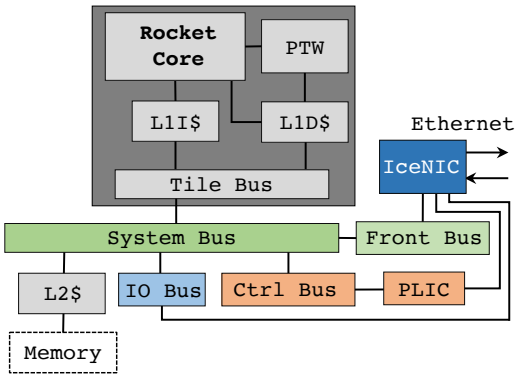
4 NANOPU PROTOTYPE

In order to evaluate the expected performance of a nanoPU, we built a prototype in Chisel [11] and C on top of the open source RISC-V Rocket Core system [7] and evaluated performance using cycle-accurate hardware simulations [64].

Figure 6a shows a high-level block diagram of the prototype's most relevant components. The prototype implements the main components of the NIC described in §3 and connects it to a slightly modified RISC-V Rocket core. The prototype implements the following aspects of the nanoPU:



(a) The nanoPU Prototype



(b) Traditional RISC-V with IceNIC

Figure 6: Block diagram of (a) the nanoPU prototype and (b) a traditional RISC-V system with IceNIC.

- The NIC packet datapath, without a hardware transport protocol, encryption/decryption or MAC processing.
- The NIC-Core Interface, including small changes to the core required to add the fast path into the register file.
- The NIC hardware thread scheduler and nanokernel.

Since the prototype does not implement a transport protocol, we assume that all messages sent and received by the applications consist of a single Ethernet packet, which we believe will be true for the vast majority of nanoservice applications.

To provide a useful reference, we compare the performance of our nanoPU prototype to a RISC-V Rocket Core with a traditional DMA-based NIC called IceNIC [40]. Figure 6b depicts a block diagram of the IceNIC RISC-V system. IceNIC behaves in a similar fashion to Intel DDIO based NICs. That is, it uses DMA to move network packets directly between the NIC and the CPU's last level cache. Since the IceNIC RISC-V design is an integrated solution, it does not include a PCIe bus–thus we expect it to exhibit lower latency than a typical modern NIC.

We made one change to the original RISC-V IceNIC design to make up for the RISC-V ISA's omission of byte-order reversal instructions. Since this operation is extremely common in network applications, and because Intel processors have hardware support for this type of operation, we added these instructions to our RISC-V processor in order to accelerate IceNIC applications. This changes allows us to obtain a more fair performance comparison with our nanoPU prototype. These instructions are not needed for nanoPU applications because the NIC swaps the byte order of message data before making it available to the application.

5 EVALUATION

We evaluated our hardware design for a nanoPU running at 3 GHz, and compared its performance to an unmodified RISC-V core with a standard NIC (IceNIC), which we will call "Traditional."

5.1 Microbenchmarks

We wrote and ran microbenchmarks for three main aspects of our nanoPU design.

Timing Analysis. We wanted to know if a nanoPU runs more slowly than the Traditional design. We synthesized both designs to run on a modern FPGA (Xilinx Ultrascale+). In both cases the critical timing path was the L2 cache, and hence our nanoPU prototype runs at the same speed as the unmodified RISC-V system.

Basic Latency/Throughput Performance Analysis. Figure 7 compares the loopback latency, as a function of packet length, for both designs. For short packets, nanoPU latency is 4× shorter than for the Traditional design. This should come as no surprise: The nanoPU transmit path is only 11 clock cycles (< 4ns at 3GHz) for a single word message, measured from when an application writes the word into the register until it is placed on the network (not including the MAC processing). And the nanoPU receive RX path is only 6 cycles (not including the MAC processing). For longer packets the loopback for the two designs eventually converges as the latency becomes dominated by store-and-forward delays.

Table 1 compares the TX and RX throughput for 64B packets for both designs; the nanoPU runs at 6–8× higher throughput than the traditional design, mostly because of the huge reductions in per-packet overheads in the nanoPU. A nanoPU application does not need to take time to instruct a DMA engine to send/receive packets, it simply reads and writes the NIC queues directly. We could expect the IceNIC throughput to increase slightly if it supported descriptor rings, which means the application doesn't need to program the DMA for every packet. However, this would also increase latency, because of batching.

	RX (Gb/s)	TX (Gb/s)
The nanoPU	116	186
RISC-V w/ IceNIC	14	28

Table 1: Comparing throughput for 64B packets on our nanoPU prototype and a traditional RISC-V core with IceNIC.

Thread Scheduler Latency. We define the latency of the thread scheduler to be the time from when an interrupt fires to the time when the first instruction of the newly selected thread is executed. We measured it to be 60 ns (180 clock cycles) on the nanoPU with almost no variation. This is at least two orders of magnitude faster than the best-known software schedulers [54]; and better still, has almost no variance, reducing the tail latency.

5.2 Bare-Metal Application Benchmarks

We implemented and evaluated four nanoservice applications, and compared their performance on our nanoPU prototype with the Traditional design.

NFV-like Streaming Application. This application emulates a very simple streaming NFV application. It treats an arriving message as a list of 8 byte unsigned integers and increments each integer then returns the resulting message back to the sender. The latency is measured from the time when the first byte of the request enters the NIC to when the last byte of the response leaves the NIC. Figure 7 shows that the nanoPU has 3.5–4× lower latency for all packet lengths. Note that the nanoPU runs almost as fast as it did for the basic loopback, whereas the Traditional design runs much slower because the CPU is forced to touch every data byte in the packet. Every byte of the packet must be copied from memory into registers, then copied back to memory and then wait for the DMA engine to send it to the NIC. The nanoPU eliminates all this overhead.

Neural Network Inference. The nanoPU core can minimize the latency required for inference by exploiting the maximum amount of model parallelism. In inference, if each node runs a small number of multiply-accumulate operations, it can be turned into a nanoservice, particularly if the weights and data are using reduced precision integers, as is typical in today's inference engines [37].

We benchmarked inference for one node in a neural network by implementing a multiply-accumulate operation. The node receives N weight messages and N data messages, one for each incoming edge, then multiplies the corresponding data and weights and accumulates the result. When all messages have been processed it sends a response message with the final result. The latency is measured from the first byte of the first incoming message to the last byte of the response.

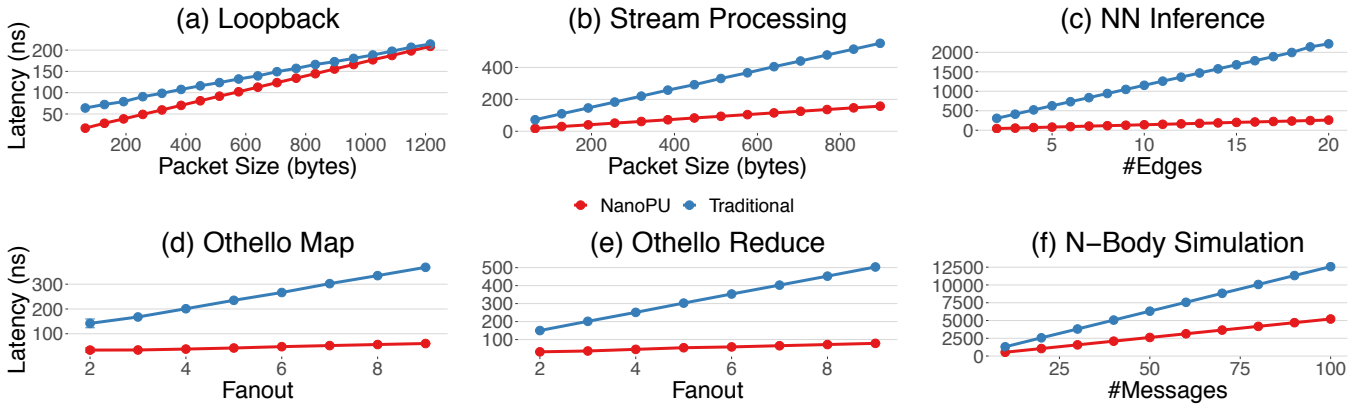


Figure 7: Nanoservice application latency comparisons for nanoPU vs a traditional RISC-V core with IceNIC.

Figure 7 shows that the nanoPU implementation has an order of magnitude lower latency than the Traditional design. This is because (1) packet data does not need to be copied from memory to registers before using the ALU, and (2) the nanoPU has a much higher RX throughput for small packets (as shown in Table 1).

Othello Map/Reduce. We re-implemented the Othello application (used as a benchmark in [31]), comparing its performance on the nanoPU with a Traditional design. An Othello board can be represented by just two 64-bit unsigned integers and hence the working set fits in the L1 cache. The set of possible next moves can be calculated in less than one microsecond, as demonstrated in [31].

We measured the latency of the "map" and "reduce" operations as a function of the *fanout*, which is the number of possible next moves given an initial board state. Because of the very short execution latency and the nanoPU's superior TX/RX throughput, the nanoPU reduces the time for both operations by a factor of 4–6× (Figure 7).

N-body Simulation. Nanoservices can efficiently implement and execute N-body simulations. This scientific computing application is typically run on HPC clusters. Astronomers use N-body simulations to model the gravitational interaction of celestial bodies. These simulations are computationally heavy, with most time spent computing the gravitational force each body exerts on every other body. A popular data structure for N-body simulations is a quad-tree, as described in the Barnes-Hut algorithm [27]. This algorithm reduces the computational complexity required for the gravity computation step from $O(N^2)$ to $O(N \log N)$.

A naïve nanoservices implementation of the Barnes-Hut algorithm would simply implement each node of the quad-tree as a separate nanotask and pass messages between nanotasks to compute gravitational forces, as shown in Figure 8. Leaf nodes represent bodies being simulated and internal nodes represent regions of space. To compute the force exerted on body F, the root node will send a message intended to traverse the quadtree. If the center of mass of the receiving node is sufficiently far away, the current node will compute the force that it exerts on body F and return the result in a response message. Otherwise, the node will forward the request to all of its children which recursively implement the same procedure.

In the Barnes-Hut algorithm, The state per node would be L1 cache resident, consisting primarily of the node's position and mass. A node computes the force it exerts on the body identified by the arriving message, using the equation $F_G = G\frac{m_1 \cdot m_2}{r^2}$ where, G is the gravitational constant, m_1 and m_2 are the masses of the two bodies, and r is the distance between them. Thus, the computation only requires a few floating point operations and the response time for each message is easily $< 1\mu$s. This is a radical departure from how the Barnes-Hut algorithm is implemented today in Changa [33], in which the quad-tree is stored as a large in-memory data structure distributed across many machines.

We implemented and measured a single node of an N-body physical simulation. Latency is measured from when the first byte of the first request message enters the NIC to when the last byte of the last response message leaves the NIC. As shown in Figure 7, the nanoPU reduces latency by 2.5× compared to a Traditional design. Section 5.4 describes a large-scale evaluation of the N-body physical simulation across multiple nodes.

5.3 Thread Scheduling Benchmarks

We compare two thread scheduling policies. The first is our NIC-driven hardware thread scheduling policy described in Section 3.3 in which the NIC decides which thread to run next and tells the CPU via the CSR registers. The second represents a more traditional Linux-style, timer-driven thread

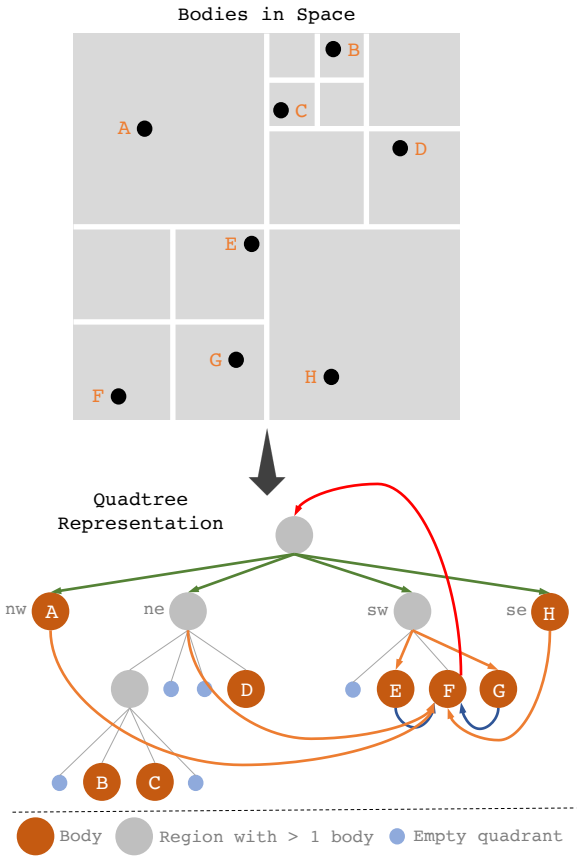


Figure 8: An example message-passing pattern to compute the gravitational forces exerted on body F using a simple nanoservices implementation of the Barnes-Hut algorithm [27].

scheduler in which timer interrupts are configured to fire periodically to decide which thread to run next. NIC hardware interrupts are disabled. In both cases, when an interrupt fires, the processor traps into the nanokernel, which then swaps to the highest priority context with messages to process. The key difference is that in the timer-driven approach, context switches only happen on periodic time boundaries and hence we can expect them to have longer latency (average and latency).

To compare the two scheduling policies we pin one high-priority context and one low-priority context on the core, then send in a stream of randomly interleaved high- and low-priority requests (separated by an interpacket gap of 500 ns). Each application services the request for 500 ns (1500 instructions) before sending back a response.

Figure 9 compares the two scheduling policies by measuring the distribution of response times for high-priority messages. Our hardware NIC-driven scheduling policy reduces the tail response latency by a factor of 5.5× and the

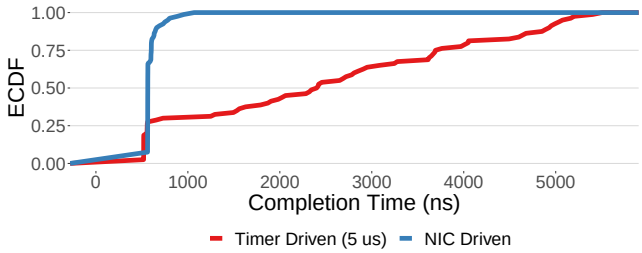


Figure 9: Latency reduction as a result of NIC-driven thread scheduling relative to timer-driven thread scheduling.

standard deviation of the response times by almost 20×. This result should not be surprising because the NIC-driven scheduling policy is able to immediately evict the low-priority thread once a high-priority message arrives, whereas the timer-driven policy must wait for a timer interrupt. The NIC-driven scheduling policy also accelerates the low-priority application because the timer-driven policy causes many traps into the nanokernel for no particular reasons, degrading its throughput. Our evaluation demonstrates the huge potential benefit of allowing the NIC hardware to drive the context-switch logic.

5.4 Large-Scale N-Body Simulation

We demonstrate the huge performance gains of a large-scale nanoservice deployment using N-body simulation and the naïve nanoservice implementation, described in §5.2. In this approach the root node of the quad-tree becomes the bottleneck, because it must process one message from each body being simulated, and the root node dictates the total runtime of the gravity computation step. We compare the performance of our naïve nanoservice application with the performance obtained when using ChanGa [33], a popular N-body gravitational simulator. Table 2 compares the average gravity computation time for 80K bodies using both the ChanGa simulator (with recommended parameter settings) and the naïve nanoservices implementation. We obtain the expected nanoservice performance by extrapolating the performance of the nanoPU bare-metal N-body node, evaluated in §5.2, to 80K messages, mimicking the processing required by the root node of the quad-tree. Even our simple, naïve nanoservice implementation will reduce the total gravity computation time by three orders of magnitude. We can drive it down further by replicating certain nodes of the quad-tree. Although ChanGa is widely used, we can significantly outperform it because, unlike nanoservices, it uses very coarse-grain parallelism. Note that our evaluation assumes the transport protocol is able to handle an 80K-to-1 incast without ever overflowing or underflowing the receiver buffer. This type of massive incast is not special to this one application, but we

	Avg. Gravity Computation Time
Nanoservices	4 ms
ChanGa	30,000 ms

Table 2: Comparing the average gravity computation time in an N-body simulation of 80K bodies using a theoretical nanoservices deployment and a real implementation using ChanGa [33].

expect that this will be common across many nanoservice applications. Designing or adopting transport protocols to deal with this degree of incast may be challenging and is a subject of future research.

6 REPRODUCIBILITY

Our nanoPU prototype and programs are public domain and open-source. We will make them publicly available with the published paper. We will provide a virtual machine with all tools and data pre-installed to make it easy for others to reproduce or extend our results.

7 DESIGN CONSIDERATIONS

Our nanoPU is deliberately made simple, in part because of the constraints of an academic research project. The designer of a complete nanoPU will need to consider many additional aspects of the design. We address some of the key tradeoffs here.

GPRs vs CSRs. Our nanoPU prototype on the RISC-V Rocket core repurposes two of its 32 general-purpose registers (GPRs) for the head and tail of the network queues. In a CPU with 32 registers, we can likely afford to lose two; our benchmark nanoservice applications did not suffer from the loss of two GPRs. However, if they did, the cost would be high as variables would spill over into memory, increasing latency. This would be exacerbated in embedded applications for a CPU with fewer GPRs. We therefore experimented with a design where the head and tail registers are implemented using control status registers (CSRs) instead, which might need to be considered in settings with limited GPRs.

Maintaining order. Our nanoPU prototype is built on top of the simple 5-stage, in-order RISC-V Rocket core. While our prototype required very minor modifications to the CPU pipeline, the changes might be more invasive on an out-of-order RISC processor in order to ensure that words are read from the RX queue in the correct order. This is even evident on our simple 5-stage core: if there is a branch misprediction when reading the network RX queue, the CPU pipeline's decode stage causes a state change which must be undone when the pipeline is flushed. We modified the RX queue to "unread" the last two words that were removed.

Floating point. Our nanoPU design minimizes the latency from the network into the *integer* register file. Many scientific computing applications, such as the N-body simulation, rely on *floating point* arithmetic, with their own set of floating point GPRs. The default RISC-V ISA does not include instructions to copy an integer register to a floating point register; the value must first be stored into memory and then loaded into the floating point register file, increasing latency. If necessary, a designer should add instructions to the RISC-V ISA to copy between the two types of GPR.

In-order execution. The nanoPU prototype is our first step towards building a domain specific-processor for nanoservice applications requiring extremely fine-grained parallelism. So far, we focused on, what we believe to be, the critical latency path: the network interface and thread scheduler, along with the accompanying changes to the compiler and assembler. A more complete design should explore the topic of out-of-order, speculative execution, common in modern CPUs. Network applications written for the nanoPU must process message data in FIFO order.

Domain specific versus general purpose. For clarity, we describe nanoPU here as if it only runs nanoservices. This might be ideal, particularly if many applications can exploit its low-latency design. We can imagine a single chip with hundreds or thousands of nanoPU cores, with little or no cache memory. But in other settings, for example in a warehouse computer, it might be necessary (or economically prudent) to support general purpose workloads too. As-is, nanoPU does not support *memory-intensive* distributed applications, such as a Key-Value store [49] (unless the database can be sharded so small that it fits in the L1 cache). In this case, the nanoPU optimizations might be added to an otherwise general-purpose CPU. The NIC, for example, could steer only nanotask messages to the GPRs, and steer memory intensive traffic (e.g. RDMA or regular TCP) to the traditional DMA, memory, and cache path. This would require a more sophisticated thread scheduler for very low latency.

Domain specific languages. We assume that the NIC pipeline is programmed using the P4 language, or similar, for fast header processing. It is worth considering whether C is the correct language for nanotasks: Perhaps a set of libraries, restricted C, or a special language could help programmers craft very fast nanotasks using explicit message processing.

OS and scheduling. Our prototype replaces Linux with our own minimal operating system, the nanokernel, designed to minimize average and peak latency for context switches. Our nanokernel currently lacks virtual memory and multiple privilege modes; applications and the nanokernel currently execute in the same address space and at the same privilege level. Our next generation nanokernel will address these issues, although we do not anticipate much increase in latency.

A real system might also need a runtime resource scheduler to ensure high utilization of nanoPU cores and NICs by dispatching nanoservice apps to nanoPU cores, monitoring the utilization of the cores and NICs, and the status of contexts, and rebalancing over/under-utilized nanoPUs on a longer timescale.

Nick: to work on para below.... incorp with privilege modes.

Network to register file. The J-machine [19] first used GPRs for extremely low-latency communication between cores. The approach was abandoned because isolating multiple contexts running on the same core required too much hardware. With many generations of improvement in process technology (from 1M transistors per chip in 1989 to 20Bn today), the per-context queues in the nanoPU solve this problem in a manner not feasible at the time of the J-machine.

8 RELATED WORK

Building Distributed Applications. Microservices and serverless compute is growing in popularity for finer-grained computing than in the past [8, 10, 26], and has been used to implement fine grained video encoding [23], compilers [22], and scientific computing applications [36]. Our nanoservices framework would drive this approach to the extreme and the nanoPU is designed to make it possible.

Minimizing RPC Tail Latency. Several authors have recently described novel approaches to minimize tail RPC completion time (Shinjuku [38], Shenango [54], and RPC-Valet [18]) by efficiently load balancing incoming requests across cores. While demonstrably good for software *micro*services, software schedulers are too slow for nanoservices. We believe load balancing nanotasks is best done by the sender or by the network, not the receiving host where it is already too late, especially for the massive degrees of incast that we expect nanoservices to exhibit. Fine-grained load-balancing across nanoPU, plus receiver-driven scheduling in the transport layer (a la Homa [47] and NDP [28]) should balance load better than a software scheduler could at the receiver.

Global Clock Synchronization. If nanoPUs have tightly synchronized clocks (within say 100ns), like SIMON [24] or PTP [62], the sender can make informed load balancing decisions, or delay sending an RPC to avoid queues and drops. This is now much easier to deploy with the advent of per-packet INT [41] with nanosecond resolution to eliminate noisy synchronization signals caused by queueing delay.

NIC Hardware. Smart NICs (NICs with CPUs on them) are becoming more popular, particularly for shared cloud data centers [9, 43, 48]. They aim to offload infrastructure software from the revenue-generating host CPUs onto smaller, lower power CPUs on the NIC. These designs are likely to increase latency for nanoservices, unless they contain a dedicated nanoPU-like fastpath for RPC messages on the NIC's data path. Alternatively, the CPUs on the NIC could include nanoPUs, to provide low latency RPCs without bothering the main host CPU.

Integrated NIC. We are not the first to propose an integrated network interface. Scale-out NUMA [50] accelerates RDMA-style applications by integrating the NIC into the machine's local cache-coherence hierarchy. The recent Compute Express Link (CXL) [17] standard is similar; it maintains coherence between the CPU memory and the memory on PCIe-attached devices. These approaches, while good for throughput, are not optimal for nanoservices, which require minimal latency into the CPU, not memory.

9 CONCLUSION

We believe that compute-intensive distributed applications would be designed very differently than they are today if utilizing remote computation is predictably only slightly slower than local computation. We believe that these applications would be implemented in an extremely fine grained manner – using the nanoservices framework. The nanoPU described in this paper is the first step towards making nanoservices a reality. There is still much work to be done. For example, designing an optimal transport protocol for nanoservices, rethinking the nanoPU's memory architecture, as well as potentially its instruction set architecture (ISA).

REFERENCES

[1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, 2015.

[2] Adiga, N. R., Almási, G., Almasi, G. S., Aridor, Y., Barik, R., Beece, D., Bellofatto, R., Bhanot, G., Bickford, R., Blumrich, M., et al. An overview of the BlueGene/L supercomputer. In *SC'02: Proceedings of the 2002 ACM/IEEE Conference on Supercomputing* (2002), IEEE, pp. 60–60.

[3] Ajayi, T., Al-Hawaj, K., Amarnath, A., Dai, S., Davidson, S., Gao, P., Liu, G., Lotfi, A., Puscar, J., Rao, A., et al. Celerity: An open source risc-v tiered accelerator fabric. In *Symp. on High Performance Chips (Hot Chips)* (2017).

[4] Alizadeh, M., Yang, S., Sharif, M., Katti, S., McKeown, N., Prabhakar, B., and Shenker, S. pFabric: Minimal near-Optimal Datacenter Transport. In *ACM SIGCOMM* (2013).

[5] Ao, L., Izhikevich, L., Voelker, G. M., and Porter, G. Sprocket: A Serverless Video Processing Framework. In *ACM SoCC* (2018).

[6] Apache Hadoop. https://hadoop.apache.org/. Accessed on 02/03/2020.

[7] Asanovic, K., Avizienis, R., Bachrach, J., Beamer, S., Biancolin, D., Celio, C., Cook, H., Dabbelt, D., Hauser, J., Izraelevitz, A., et al.

The rocket chip generator. *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2016-17* (2016).

[8] AWS Lambda. https://aws.amazon.com/lambda/. Accessed on 06/28/2019.

[9] Aws nitro system. https://aws.amazon.com/ec2/nitro/. Accessed on 02/04/2020.

[10] Azure Functions. https://azure.microsoft.com/en-us/services/functions/. Accessed on 06/28/2019.

[11] Bachrach, J., Vo, H., Richards, B., Lee, Y., Waterman, A., Avižienis, R., Wawrzynek, J., and Asanović, K. Chisel: constructing hardware in a scala embedded language. In *DAC Design Automation Conference 2012* (2012), IEEE, pp. 1212–1221.

[12] Barroso, L. A., Dean, J., and Holzle, U. Web Search For a Planet: The Google Cluster Architecture. *IEEE Micro 23*, 2 (2003), 22–28.

[13] Bohnenstiehl, B., Stillmaker, A., Pimentel, J. J., Andreas, T., Liu, B., Tran, A. T., Adeagbo, E., and Baas, B. M. Kilocore: A 32-nm 1000-processor computational array. *IEEE Journal of Solid-State Circuits 52*, 4 (2017), 891–902.

[14] Bosshart, P., Daly, D., Gibb, G., Izzard, M., McKeown, N., Rexford, J., Schlesinger, C., Talayco, D., Vahdat, A., Varghese, G., et al. P4: Programming protocol-independent packet processors. *ACM SIGCOMM Computer Communication Review 44*, 3 (2014), 87–95.

[15] Bosshart, P., Gibb, G., Kim, H.-S., Varghese, G., McKeown, N., Izzard, M., Mujica, F., and Horowitz, M. Forwarding metamorphosis: Fast programmable match-action processing in hardware for sdn. *ACM SIGCOMM Computer Communication Review 43*, 4 (2013), 99–110.

[16] Carreira, J., Fonseca, P., Tumanov, A., Zhang, A., and Katz, R. Cirrus: A Serverless Framework for End-to-End ML Workflows. In *ACM SoCC* (2019).

[17] Compute express link: Breakthrough cpu-to-device interconnect. https://www.computeexpresslink.org/. Accessed on 2020-01-29.

[18] Daglis, A., Sutherland, M., and Falsafi, B. Rpcvalet: Ni-driven tail-aware balancing of μs-scale rpcs. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems* (2019), pp. 35–48.

[19] Dally, W. J., Chien, A., Fiske, S., Horwat, W., and Keen, J. The j-machine: A fine grain concurrent computer. Tech. rep., MASSACHUSETTS INST OF TECH CAMBRIDGE MICROSYSTEMS RESEARCH CENTER, 1989.

[20] Dean, J. Software engineering advice from building large-scale distributed systems. https://static.googleusercontent.com/media/research.google.com/en/us/people/jeff/stanford-295-talk.pdf. Accessed on 02/04/2020.

[21] Dean, J., and Ghemawat, S. MapReduce: Simplified Data Processing on Large Clusters. In *USENIX OSDI* (2004).

[22] Fouladi, S., Iter, D., Chatterjee, S., Kozyrakis, C., Zaharia, M., and Winstein, K. A thunk to remember: make-j1000 (and other jobs) on functions-as-a-service infrastructure, 2017.

[23] Fouladi, S., Wahby, R. S., Shacklett, B., Balasubramaniam, K. V., Zeng, W., Bhalerao, R., Sivaraman, A., Porter, G., and Winstein, K. Encoding, Fast and Slow: Low-latency Video Processing Using Thousands of Tiny Threads. In *USENIX NSDI* (2017).

[24] Geng, Y., Liu, S., Yin, Z., Naik, A., Prabhakar, B., Rosenblum, M., and Vahdat, A. SIMON: A Simple and Scalable Method for Sensing, Inference and Measurement in Data Center Networks. In *USENIX NSDI* (2019).

[25] Glassner, A. S. *An introduction to ray tracing*. Elsevier, 1989.

[26] Google Cloud Functions. https://cloud.google.com/functions/. Accessed on 06/28/2019.

[27] Grama, A. Y., Kumar, V., and Sameh, A. Scalable parallel formulations of the barnes-hut method for n-body simulations. In *Supercomputing'94: Proceedings of the 1994 ACM/IEEE Conference on Supercomputing* (1994), IEEE, pp. 439–448.

[28] Handley, M., Raiciu, C., Agache, A., Voinescu, A., Moore, A. W., Antichi, G., and Wójcik, M. Re-architecting datacenter networks and stacks for low latency and high performance. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication* (2017), pp. 29–42.

[29] Hennessy, J. L., and Patterson, D. A. *Computer Architecture, Sixth Edition: A Quantitative Approach*, 6th ed. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2017.

[30] Ibanez, S., Antichi, G., Brebner, G., and McKeown, N. Event-driven packet processing. In *Proceedings of the 18th ACM Workshop on Hot Topics in Networks* (2019), pp. 133–140.

[31] Ibanez, S., Shahbaz, M., and McKeown, N. The case for a network fast path to the cpu. In *Proceedings of the 18th ACM Workshop on Hot Topics in Networks* (2019), pp. 52–59.

[32] Intel Ethernet 800 Series NIC. https://www.intel.com/content/www/us/en/architecture-and-technology/ethernet.html. Accessed on 2020-01-29.

[33] Jetley, P., Gioachin, F., Mendes, C., Kale, L. V., and Quinn, T. Massively parallel cosmological simulations with ChaNGa. In *2008 IEEE International Symposium on Parallel and Distributed Processing* (2008), IEEE, pp. 1–12.

[34] Jin, X., Li, X., Zhang, H., Foster, N., Lee, J., Soulé, R., Kim, C., and Stoica, I. Netchain: Scale-free sub-rtt coordination. In *15th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 18)* (2018), pp. 35–49.

[35] Jin, X., Li, X., Zhang, H., Soulé, R., Lee, J., Foster, N., Kim, C., and Stoica, I. Netcache: Balancing key-value stores with fast in-network caching. In *Proceedings of the 26th Symposium on Operating Systems Principles* (2017), pp. 121–136.

[36] Jonas, E., Pu, Q., Venkataraman, S., Stoica, I., and Recht, B. Occupy the cloud: Distributed computing for the 99%. In *ACM SoCC* (2017).

[37] Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., Bates, S., Bhatia, S., Boden, N., Borchers, A., et al. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture* (2017), pp. 1–12.

[38] Kaffes, K., Chong, T., Humphries, J. T., Belay, A., Mazières, D., and Kozyrakis, C. Shinjuku: Preemptive scheduling for μsecond-scale tail latency. In *16th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 19)* (2019), pp. 345–360.

[39] Kalia, A., Kaminsky, M., and Andersen, D. G. Datacenter RPCs Can Be General and Fast. In *USENIX NSDI* (2019).

[40] Karandikar, S., Mao, H., Kim, D., Biancolin, D., Amid, A., Lee, D., Pemberton, N., Amaro, E., Schmidt, C., Chopra, A., et al. Firesim: Fpga-accelerated cycle-exact scale-out system simulation in the public cloud. In *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)* (2018), IEEE, pp. 29–42.

[41] Kim, C., Sivaraman, A., Katta, N., Bas, A., Dixit, A., and Wobker, L. J. In-band network telemetry via programmable dataplanes. In *ACM SIGCOMM* (2015).

[42] Li, Y., Miao, R., Liu, H. H., Zhuang, Y., Feng, F., Tang, L., Cao, Z., Zhang, M., Kelly, F., Alizadeh, M., et al. Hpcc: high precision congestion control. In *Proceedings of the ACM Special Interest Group on Data Communication*. 2019, pp. 44–58.

[43] Mellanox bluefield-2. https://www.mellanox.com/products/bluefield2-overview. Accessed on 02/04/2020.

[44] Mellanox ConnectX-6 200GbE Ethernet Adapter IC. https://www.mellanox.com/sites/default/files/related-docs/prod_silicon/PB_ConnectX-6_EN_IC.pdf. Accessed on 2020-01-29.

[45] Miao, R., Zeng, H., Kim, C., Lee, J., and Yu, M. Silkroad: Making stateful layer-4 load balancing fast and cheap using switching asics.

In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication* (2017), pp. 15–28.

[46] Mittal, R., Lam, V. T., Dukkipati, N., Blem, E., Wassel, H., Ghobadi, M., Vahdat, A., Wang, Y., Wetherall, D., and Zats, D. Timely: Rtt-based congestion control for the datacenter. *ACM SIGCOMM Computer Communication Review 45*, 4 (2015), 537–550.

[47] Montazeri, B., Li, Y., Alizadeh, M., and Ousterhout, J. Homa: A receiver-driven low-latency transport protocol using network priorities. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication* (2018), pp. 221–235.

[48] Naples dsc-100 distributed services card. https://www.pensando.io/assets/documents/Naples_100_ProductBrief-10-2019.pdf. Accessed on 02/04/2020.

[49] Nishtala, R., Fugal, H., Grimm, S., Kwiatkowski, M., Lee, H., Li, H. C., McElroy, R., Paleczny, M., Peek, D., Saab, P., et al. Scaling memcache at facebook. In *Presented as part of the 10th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 13)* (2013), pp. 385–398.

[50] Novakovic, S., Daglis, A., Bugnion, E., Falsafi, B., and Grot, B. Scale-out numa. *ACM SIGPLAN Notices 49*, 4 (2014), 3–18.

[51] Nvidia: GeForce Now. https://www.nvidia.com/en-us/geforce-now/. Accessed on 02/04/2020.

[52] OpenFaaS: Serverless Functions, Made Simple. https://www.openfaas.com/. Accessed on 02/03/2020.

[53] Options for Code Generation Conventions. https://gcc.gnu.org/onlinedocs/gcc/Code-Gen-Options.html#Code-Gen-Options. Accessed on 02/04/2020.

[54] Ousterhout, A., Fried, J., Behrens, J., Belay, A., and Balakrishnan, H. Shenango: Achieving high {CPU} efficiency for latency-sensitive datacenter workloads. In *16th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 19)* (2019), pp. 361–378.

[55] PerfKit: gRPC Performance. https://performance-dot-grpc-testing.appspot.com/explore?dashboard=5652536396611584&widget=1747498211&container=897341821. Accessed on 02/03/2020.

[56] Rapaport, D. C., and Rapaport, D. C. R. *The art of molecular dynamics simulation.* Cambridge university press, 2004.

[57] RapidIO Intel FPGA IP User Guide. https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/ug/ug_rapidio.pdf. Accessed on 02/04/2020.

[58] Ricart-Sanchez, R., Malagon, P., Alcaraz-Calero, J. M., and Wang, Q. Hardware-accelerated firewall for 5g mobile networks. In *2018 IEEE 26th International Conference on Network Protocols (ICNP)* (2018), IEEE, pp. 446–447.

[59] Risc-v. https://riscv.org/. Accessed on 2020-01-29.

[60] Schuehler, D. V., and Lockwood, J. W. A Modular System for FPGA-based TCP Flow Processing in High-Speed networks. In *FPL* (2004), Springer.

[61] Serverless computing in financial services: High performance computing—with more speed, accuracy and flexibility. https://www.ibm.com/downloads/cas/Q93RMQMN. Accessed on 02/03/2020.

[62] Shankarkumar, V., Montini, L., Frost, T., and Dowd, G. Precision Time Protocol Version 2 (PTPv2) Management Information Base. RFC 8173, June 2017.

[63] Tofino. https://www.barefootnetworks.com/products/brief-tofino/. Accessed on 02/04/2020.

[64] Verilator. https://www.veripool.org/wiki/verilator. Accessed on 2020-01-29.

[65] Zhang, W. *State-space Search: Algorithms, Complexity, Extensions, and Applications.* Springer Science & Business Media, 1999.

[66] Zhu, Y., Eran, H., Firestone, D., Guo, C., Lipshteyn, M., Liron, Y., Padhye, J., Raindel, S., Yahia, M. H., and Zhang, M. Congestion control for large-scale rdma deployments. *ACM SIGCOMM Computer Communication Review 45*, 4 (2015), 523–536.