Adva Simantov Damti

From: Parkhurst, Jeff <jeff.parkhurst@intel.com>
Sent: Wednesday, 6 July 2022 21:41
Cc: Parkhurst, Jeff
Subject: Call for proposals
Attachments: Arch SRS Proposal Template _PI.docx; Arch SRS RFP_PI.pptx

Follow Up Flag: Follow up
Flag Status: Flagged

Dear Professor/administrator,

Please consider the call for proposals below sent to you based upon your field of expertise or because you were identified as a contact for your school. We expect to fund selected proposals between $100K to $200K per year depending on their breadth out of a total $800K per year budget. At this time, we expect to fund research for a total of 3 years (though funding beyond 1 year is not guaranteed). The proposal collateral we are seeking is fairly lightweight and is attached. Your proposal should be no more than 4 written pages and with an accompanying slide presentation. Proposals and slides are due August 1st. We plan to have identified proposals selected by end of August. Let us know if you have any questions. Please feel free to forward to appropriate researchers at your institution as well.

Best Regards,

Frank Hady
Jeff Parkhurst

Motivation: Datacenters continue to move to heterogeneous compute with a variety of different computational engines solving specific workloads, or portions of workloads, at significantly higher performance and lower power that general purpose CPUs. This is a significant change from the CPU dominated data center, bringing with it new software, XPU and systems challenges and opportunities. For example, porting current applications from CPU to multiple XPU is challenging at the software level requiring careful partitioning across a set of XPUs and effective xPU to xPU interaction – none of which is solved in the general case. Creating an environment(HW and SW) which comprehends the variety of available domain specific accelerators in a system is required for higher performance as well as energy efficiency. The following research vectors (RVs) listed below are meant to address these challenges.

RV1: Systems Research: This vector seeks a holistic systems level approach to integration of multiple accelerators into a cohesive data center platform. Target workloads include multimedia, graphics, infrastructure processing, and graph analytics and the system defined will span HW and SW. Rather than a set of individual computational units, such a system should enable an effective and potentially fine-grained interaction between heterogeneous computational elements towards a single application. A fresh look at basic computation element interactions such as communication primitives and data sharing and the HW/SW dissection models for both is within scope. This vector should tightly integrated with x86 based processors and XPUs (Intel Arc, FPGA, Intel VPU, etc.) into the solution. Working in conjunction with RV2, we envision a highly integrated multi core compute platform allowing for ease of HW/SW co development.

RV2: SW environment and accelerator programmability: The focus of this RV is to make it easier for developers to use accelerators in conjunction with the CPU. By taking a fresh look at the programming of system heterogeneity from the outset, rather than patching homogenous systems together, a superior environment may be possible. Such a system may enable transitioning CPU or GPU centric applications to a more efficient heterogenous systems implementations without developer heroics. The mechanisms required may include software stacks, runtimes, instrumentation,

programming, and analysis tools and more potentially allowing automatic transformations to map code to the underlying platform in an optimized manner. Research may also cover definition of the basic software constructs needed for specifying and managing data and computations handoff.

RV3: IPU system integration: The integration of Infrastructure Processors (IPUs) offers a potential path to more efficient heterogenous system. Definitional Infrastructure Processor research is needed to cover more natural integration with the rest of the heterogenous system. IPU and CPU co-design for microservices exploring system wide workload design for offload to IPU in light of changing memory hierarchies may provide significant advantage. More efficient thread scheduling, custom instructions, interrupt handling & event processing, operating system advances all in the context of both internal IPU operation and external interaction maybe fruitful. Finally, research on programmable IPU protocol engines and finally tools and methods for verifying the correctness of network transport protocol implementations against a formal specification is also of interest.