Hal Finkel

Program Manager, ASCR

Presentation for ASCAC: July 29, 2021

2021 ASCR Roundtable Discussion on Operating-Systems Research

- ASCR hosted a roundtable discussion on operating-systems research on January 25, 2021.
- The roundtable discussion took place via Zoom, 12-5pm Eastern.
- Observer registration was open to the public and over 100 registered observers attended.
- Five invited speakers presented during the discussion:

Speaker	Institution
Eric Van Hensbergen	ARM
Jack Lange	Univ of Pittsburgh
Rudolf Eigenmann	Univ of Delaware
Martin Schulz	TUM (Germany)
Balazs Gerofi	RIKEN (Japan)

2021 ASCR Roundtable Discussion on Operating-Systems Research

- 16 national-laboratory scientists participated:

Participant	Laboratory
John Shalf	LBL
Costin Iancu	LBL
Barry Rountree	LLNL
Tapasya Patki	LLNL
Mike Lang	LANL
Lucho Ionkov	LANL
Ron Brightwell	SNL
Kevin Pedretti	SNL

Participant	Laboratory
Shantenu Jha	BNL
Barbara Chapman	BNL
Sriram Krishnamoorthy	PNNL
Roberto Gioiosa	PNNL
Jeff Vetter	ORNL
Christian Engelmann	ORNL
Kamil Iskra	ANL
Pete Beckman	ANL

2021 ASCR Roundtable Discussion on Operating-Systems Research

- The agenda included lab presentations, presentations by the invited speakers, and planned discussion topics:
  - Extreme Heterogeneity
  - Edge Computing
  - Data and Artificial Intelligence
  - Containers, Virtualization, and Programming-Environment Management
  - Resiliency and Correctness

- Two whitepapers were prepared to capture the discussions at the roundtable:
  - *Research Opportunities in Operating Systems for High-Performance Scientific Computing*
  - *Research Opportunities in Operating Systems for Scientific Edge Computing*

  - Both whitepapers have been posted along with this presentation on the ASCAC website.

  - What follows is a presentation of the whitepapers; the views expressed are those of the whitepaper authors.

Research Opportunities in Operating Systems for High-Performance Scientific Computing

Contributors: Pete Beckman (ANL), Ron Brightwell (SNL), Rudi Eigenmann (Univ of Delaware), Christian Engelmann (ORNL), Roberto Gioiosa (PNNL), Kamil Iskra (ANL), Shantenu Jha (BNL), Jack Lange (Univ of Pittsburgh), Tapasya Patki (LLNL), Kevin Pedretti (SNL)

DOE Point of Contact: Hal Finkel <hal.finkel@science.doe.gov>

May 18, 2021

- The primary aspects of the modern computing ecosystem key to understanding future research opportunities in operating systems are:
  - Full-stack co-design for extreme heterogeneity and scalability
  - Adaptive management and partitioning of resources
  - Smart supercomputer systems and facilities

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for High-Performance Scientific Computing*.

- Full-stack co-design for extreme heterogeneity and scalability:
  - Current trends:
    - Increasing heterogeneity, increasing system scales, and increasing data migration;
    - Software must increasingly be optimized and specialized for different kinds of hardware;
    - CPUs, computational accelerators (e.g., GPUs), programming network interfaces, etc., all have programming environments requiring specialization at different levels.

    - Systems research and development increasingly requires co-design of both the hardware and the software.
    - With high-quality open-source hardware designs and fast cycle-accurate simulation tools, it is now possible to rapidly co-design pre-silicon hardware together with full software stacks.
    - For example, this capability was recently used to port KVM [the Linux *Kernel-based Virtual Machine*] to the RISC-V architecture prior to final ratification of the RISC-V virtualization extensions, providing critical validation and feedback.

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for High-Performance Scientific Computing*.

- Full-stack co-design for extreme heterogeneity and scalability:
  - Research Challenges and Open Questions:
    - Proving out new OS ideas is difficult due to researchers having
      - Limited ability to affect hardware designs
      - Limited access to large scale HPC platforms on which custom OS software can be booted
    - Performing experiments via FPGA-accelerated simulation tools attacks both challenges.

    - Operating-systems research in the context of full-stack co-design can address:
      - Rapidly developing novel accelerator hardware and accompanying software support
      - Common interfaces to heterogeneous hardware enabling effective OS management
      - Autonomous resource management that reduces the burden on users
      - System-wide security that enables breaking free from the “root is special” model

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for High-Performance Scientific Computing*.

- Full-stack co-design for extreme heterogeneity and scalability:
  - Research Challenges and Open Questions:
    - Specific research questions include:
      - How should we think about the OS managing extremely heterogeneous systems, and what new hardware and software mechanisms are needed to enable this?
      - What are the right interfaces to communicate the needed information between programs, compilers, runtime/OS, monitoring subsystem architecture?
      - Can autonomous resource management, assisted by co-designed hardware and OS software, reduce the user burden for managing extremely heterogeneous devices and memories?
      - Can this be done portably, such that difficult porting efforts are not required when moving from one system to the next?
      - How to design an intelligent memory management featuring automated, on-the-fly data optimization (compression, transformations, placement)?

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for High-Performance Scientific Computing*.

- Full-stack co-design for extreme heterogeneity and scalability:
  - Research Challenges and Open Questions:
    - Specific research questions include: [continued]
      - What role can the OS play in identifying inefficient resource usage and suggesting or automating potential improvements?
      - What new hardware capabilities could help the OS do this?
      - Through hardware and software co-design, can we enable system-wide security such that root is no longer a special thing on HPC systems, similar to the case on public clouds?
      - Would this reduce the facility burden for managing security issues, as well as enable new types of cloud-native workloads to run on DOE supercomputers?

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for High-Performance Scientific Computing*.

- Full-stack co-design for extreme heterogeneity and scalability:
  - Benefits of Success:
    - As rapid, full-stack co-design continues to mature, it may become feasible to rapidly develop and economically deploy domain-optimized silicon targeted at DOE computing workloads, delivering world-class efficiency and performance.
    - Success would produce highly credible OS functionality, hardware designs, and actionable information which could be used to more effectively influence our vendor partners, ultimately leading to higher performing and more efficient DOE supercomputer platforms.

  - Contributing Research Communities:
    - OS research communities;
    - A vibrant open-source hardware community in academia;
    - Leverages a new class of hardware vendors that are emerging to help customers develop domain optimized hardware targeting their particular workloads.

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for High-Performance Scientific Computing*.

- Adaptive management and partitioning of resources:
  - Current trends:
    - An increasing number of extreme-scale applications are composed of multiple components, workflows that must often be “rewired” dynamically.
    - The price of static decisions about workload placement and execution, and their management increases for these workloads with scale and heterogeneity.
    - As the computational capabilities of HPC compute nodes keep increasing, yet fully leveraging those capabilities using individual applications gets ever harder.
    - Running a diverse mix of workloads on a node would improve the overall resource utilization, however this is often avoided, especially the multi-user scenarios, because this tends to introduce performance variability.

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for High-Performance Scientific Computing*.

- Adaptive management and partitioning of resources:
  - Current trends: [continued]
    - A lack of standards currently hampers the construction of composable, extensible scientific workflows.
    - Currently, nearly every multi-component application and workflow assumes a different model of resource partitioning, selection, and availability, as well as different interfaces, resource state models, and resource / process management.
    - This makes applications and workflows brittle and the barrier to portability and extensibility significant.

    - Virtualization techniques are a popular existing solution in cloud environments, and offer partial solutions to some of the adaptive management/partitioning challenges, but they have traditionally been avoided in HPC due to overheads.
    - HPC-specific multi-kernels can offer lower overheads at the cost of additional implementation complexity.

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for High-Performance Scientific Computing*.

- Adaptive management and partitioning of resources:
  - Current trends: [continued]
    - Adoption of containers is increasing, also offering partial solutions to some management and partitioning challenges, but bringing new challenges around composability and efficient workflow integration.

    - Extreme-scale applications are already facing limited memory capacity and bandwidth, yet there are no widely accepted, standardized interfaces for exploiting deep memory hierarchies.
    - There is an inherent conflict between how much of the memory hierarchy to hide from users (so as not to overly complicate the programming) vs how many novel capabilities to expose (to take advantage of the possible performance improvements).

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for High-Performance Scientific Computing*.

- Adaptive management and partitioning of resources:
  - Research Challenges and Open Questions:
    - Open questions regarding how applications, programming environments, and systems can exchange information and cooperatively react to changing workloads and resources include:
      - How to extend programming languages and compilers so that users can express the user-relevant information?
      - What is the communication architecture (i.e., what information is created, communicated, and consumed where and when)?
      - How can current programming methods and compiler optimizations be extended to take advantage of the new information?
      - How can dynamic optimization systems be constructed that adapt at runtime and in the field to changing environments?
      - How to build performance models and decision support to guide users and compilers in these methods and optimizations?

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for High-Performance Scientific Computing*.

- Adaptive management and partitioning of resources:
  - Research Challenges and Open Questions:
    - Open research questions include: [continued]
      - How to construct data-centric abstractions that are application-aware and that present a range of interfaces for a right fit to every application and runtime system, from fully transparent to fully explicit ones?
      - How to enable integration across all of the on-node devices, including function offloading to the NIC?
      - How to reintroduce resource partitioning capabilities in HPC environments, overcoming concerns regarding security and reproducibility?
      - How to support containers within an HPC environment, including workflow integration and performance monitoring?
      - What is the performance price of static execution and resource management as a function of scale and heterogeneity?
      - What are “intrinsic” scales of resource partitioning as a function of scale and dynamism?

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for High-Performance Scientific Computing*.

- Adaptive management and partitioning of resources:
  - Research Challenges and Open Questions:
    - Open research questions include: [continued]
      - In the presence of adaptive execution and dynamic resource behavior, what is the trade-off between global versus partitioned resource management at exascale?
      - What are the challenges of providing the application complete control of the resources?
      - How is information propagated across resources partitions?
      - How can adaptive execution be implemented without incurring significant overhead?

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for High-Performance Scientific Computing*.

- Adaptive management and partitioning of resources:
  - Benefits of Success:
    - Middleware capabilities that
      - Provide the abstractions and interfaces for distributed and high-performance resource management
      - Allow a broad range of multi-component applications and workflows that are agnostic to the specifics of underlying resources and platforms
      - Make it easier to scale up and out.
    - Significant performance improvements
    - Improved resource utilization (less movement, better power consumption)
    - Improved productivity for users/developers when porting to new platforms
    - Resource-partitioning improvements will result in
      - Better utilization of available node resources
      - Better performance for dynamic workloads (including ensemble calculations and AI)
      - Opening HPC systems to a broader, more diverse set of workloads

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for High-Performance Scientific Computing*.

- Adaptive management and partitioning of resources:
  - Contributing Research Communities
    - Research communities across the entire system stack will need to collaborate
    - DOE applications, facilities, and systems researchers
    - Software Engineering for Adaptive and Self-managing Systems (see: https://conf.researchr.org/home/seams-2021)
    - Middleware design principles, programming abstractions and paradigms for reconfigurable, adaptable, and reflective approaches (see: ACM Middleware).
    - Synergies exist with the cloud and edge communities.

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for High-Performance Scientific Computing*.

- Smart supercomputer systems and facilities:
  - Current trends:
    - Operational intelligence (OI) optimizes the efficiency and effectiveness of systems and facilities using an observe-orient-decide-act loop for adaptation.
    - The loop consists of operational data aggregation, operational data analytics, decision making that considers trade-offs, and operational configuration actions.
    - Corrective actions are often highly limited, as the resilience "toolbox", i.e., the number and types of corrective actions, employed in today's computing systems and facilities is highly limited.

    - DOE's recent Computational Facilities Research Workshop report identified smart systems and facilities as a broad challenge area with enabling automation and eliminating human-in-the-loop requirements as a cross-cutting theme.

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for High-Performance Scientific Computing*.

- Smart supercomputer systems and facilities:
  - Research Challenges and Open Questions:
    - Specific research challenges include:
      - Autonomous resource management at different granularities: programming model runtime, node OS, global OS and facility
      - Machine-in-the-loop feedback through operational intelligence: monitoring, operational data analytics, autonomous decision making and adaptive resource management
      - Improving operational productivity and lowering operational costs through corrective actions
      - Autonomous adaptation to system properties and application needs
      - Identification of relevant monitoring data for specific control problems
      - Understanding and modeling the involved trade-offs
      - Leveraging community and industry software for reuse and maintainability
      - Enabling real-time control

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for High-Performance Scientific Computing*.

Research Opportunities in Operating Systems for Scientific Edge Computing

Contributors: Pete Beckman (ANL), Christian Engelmann (ORNL), Shantenu Jha (BNL), Jack Lange (Univ of Pittsburgh)

DOE Point of Contact: Hal Finkel <hal.finkel@science.doe.gov>

May 18, 2021

- As scientific experiments generate ever-increasing amounts of data, and grow in operational complexity, modern experimental science demands unprecedented computational capabilities at the edge -- physically proximate to each experiment.
- Scientific edge computing has a number of unique challenges

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for Scientific Edge Computing*.

- Current trends:
  - Edge computing is set to become an important feature in future HPC environments as scientific workloads begin to incorporate live data collection and analysis.
  - Examples of this trend include:
    - Streaming data directly from large scale scientific instruments and experimental installations to high end computing environments
    - Large scale data collection and analysis needed by distributed sensor platforms
    - Real time computational steering of “smart” infrastructure including experimental and manufacturing processes
  - The current state of practice in experimental science heavily involves human-in-the-loop activity for controlling instruments and computing resources, analyzing data, steering ongoing experiments and planning future studies.
  - In several cases, the data stream is too complex for online control by humans, necessitating multiple runs with different parameter settings and neglecting online control.
  - This results in reduced precision of experiments, particularly in materials synthesis applications.

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for Scientific Edge Computing*.

- Current trends: [continued]
  - Current human-in-the-loop solutions for steering instrument experiments with computational analysis have been developed on an ad-hoc basis for specific instruments and are limited in scope.
  - Data formats, communication protocols and human-machine interfaces are usually incompatible with other instruments, analysis approaches and the envisioned autonomous online real-time control.

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for Scientific Edge Computing*.

- Research Challenges and Requirements
  - Incorporating large numbers of edge resources will require scalable access control mechanisms that are able to support much more heterogeneous computing resources and more complex organizational structures.
  - To effectively adapt to these environments will require access control mechanisms that support a high degree of semantic expressiveness in order to effectively encode increasingly complex and abstract policies.
  - Access control policies will need to expand to include heterogeneous and dynamic collections of edge-based hardware resources which will likely exhibit large amounts of churn and varying degrees of availability.
  - This will require the expansion of security models to enable them to dynamically establish trust with potentially ephemeral edge resources in order to ensure end-to-end data security and access control.
  - Trusted edge resources will need to support the ability to provide trusted execution environments with hardware level primitives as well as attestable and authenticatable software environments.

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for Scientific Edge Computing*.

- Research Challenges and Requirements [continued]
  - The recent AI for Science Report outlines the need for smart systems, instruments and facilities to enable science breakthroughs with autonomous experiments, “self-driving” laboratories, smart manufacturing, and AI-driven design, discovery and evaluation.
  - DOE’s recent Computational Facilities Research Workshop report identified smart systems and facilities as a broad challenge area with enabling automation and eliminating human-in-the-loop requirements as a cross-cutting theme.

  - One of the major bottlenecks for science is the limited speed at which experiments can receive feedback from computation and theory.
  - Accelerating this cycle requires connected instrumentation with local/instantaneous computation using edge computing resources where feasible and with remote/urgent computation using leadership computing resources when necessary.
  - Motivating use cases: a biological sample may have only a limited lifetime, real-time feedback to discover and treat manufacturing defects

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for Scientific Edge Computing*.

- Research Challenges and Requirements [continued]
  - OS (middleware) for scientific edge-computing must include agile mechanisms to provision distributed resources with collective properties and possibly inconsistent capabilities and availability.
  - Top down and centralized resource federation is unlikely to scale, be resilient, or even responsive to dynamic changes.
  - In order to achieve the necessary control and information flow for integrated application-systems resource and workload management, the infrastructure must also present unified abstractions and interfaces for resource and workload management to distributed applications.
  - For example, user-facing abstractions and interfaces must include explicit performance and quality of service measures; system-facing abstractions and interfaces must include information about resource availability, and control for configuration and distributed resources selection.

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for Scientific Edge Computing*.

- Open Questions:
  - Areas of open questions include:
    - Distributed hardware and software architectures
    - Communication protocols and interfaces
    - Resource orchestration and control
    - Data transfer, management and provenance
    - Leveraging AI/ML, data analytics and advanced statistics for computational steering and planning of experiments
    - The efficient and effective integration of human-machine interfaces
    - The role of virtual instruments as digital twins

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for Scientific Edge Computing*.

- Benefits of Success
  - Will allow future HPC systems to dramatically expand their utility by supporting emerging classes of workloads based on large scale and distributed online data collection.
  - Enabling [while maintaining trust] autonomous experiments, “self-driving” laboratories, smart manufacturing, and AI-driven design, discovery and evaluation using a combination of edge and center computing and data resources enable faster science breakthroughs with autonomous steering of ongoing and planning of the next experiments.

- Contributing Research Communities
  - The edge computing community
  - The distributed systems community
  - The HPC global OS community
  - HPC operations personnel at computing facilities
  - The AI/ML and data analytics community
  - The decision sciences community
  - The instrument science community

Views expressed are those of the authors of the whitepaper *Research Opportunities in Operating Systems for Scientific Edge Computing*.