LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-831695

Non-Proprietary Companion to the Q1 CY2021 PathForward Final Assessment WBS 2.4.1, Milestone PM-HI-1040

B. R. de Supinski, S. Atchley, C. Hughes, R. Goldstone, H. Finkel, I. Karlin, S. Pakin, C. Daley

February 10, 2022

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

EXASCALE COMPUTING PROJECT

ECP-U-HI-RPT_2021_00144

Non-Proprietary Companion to the Q1 CY2021 PathForward Final Assessment WBS 2.4.1, Milestone PM-HI-1040

Bronis R. de Supinski, LLNL
L3 Lead for ECP PathForward
Scott Atchley, ORNL
ECP TR for AMD
Clay Hughes, SNL
ECP TR for Cray, Inc.
Robin Goldstone, LLNL
ECP TR for HPE
Hal Finkel, ANL, and
Ian Karlin, LLNL
ECP TR for Intel
Scott Pakin, LANL
ECP TR for IBM
Christopher Daley, LBNL
ECP TR for NVIDIA

March 8, 2021

U.S. DEPARTMENT OF ENERGY | Office of Science

National Nuclear Security Administration

ECP-U-HI-RPT_2021_00144

Non-Proprietary Companion to the ECP Milestone Report
Q1 CY2021 PathForward Final Assessment
WBS 2.4.1, Milestone PM-HI-1040

Office of Advanced Scientific Computing Research
Office of Science
US Department of Energy

Office of Advanced Simulation and Computing
National Nuclear Security Administration
US Department of Energy

March 8, 2021

ECP Milestone Report
Q1 CY2021 PathForward Final Assessment
WBS 2.4.1, Milestone PM-HI-1040

APPROVALS

Submitted by:

Bronis R. de Supinski, Lawrence Livermore National Laboratory — Date
L3 CAM for PathForward
Exascale Computing Project

Concurrence:

Doug Kothe, Oak Ridge National Laboratory — Date
Project Director
Exascale Computing Project

Lori Diachin, Lawrence Livermore National Laboratory — Date
Deputy Project Director
Exascale Computing Project

Katie Antypas, Lawrence Berkeley National Laboratory — Date
Director – Hardware and Integration
Exascale Computing Project

Susan Coghlan, Argonne National Laboratory — Date
Deputy Director – Hardware and Integration
Exascale Computing Project

Scott Atchley, Oak Ridge National Laboratory — Date
TR for AMD
Exascale Computing Project

Clay Hughes, Sandia National Laboratories — Date
TR for Cray
Exascale Computing Project

Ian Karlin, Livermore National Laboratory — Date
TR for Intel
Exascale Computing Project

Scott Pakin, Los Alamos National Laboratory — Date
TR for IBM
Exascale Computing Project

Christopher Daley, Lawrence Berkeley National Laboratory — Date
TR for NVIDIA
Exascale Computing Project

VERSION CHANGE CONTROL TABLE

Version	Creation Date	Description	Approval Date
1.0	1/25/2021	Derived from proprietary report draft	
1.1	3/8/2021	Incorporated TR feedback	
2.0	10/13/2021	Updated for final contract milestone status	

CONTENTS

ACRONYMS

AI	Artificial Intelligence
AMD	Advanced Micro Devices
APP	Accelerated Processing Platform
CAM	Control Account Manager
CMI	Component and Memory Integration
COG	Compute Optimized GPUs
CORAL	Collaboration of Oak Ridge, Argonne and Livermore
DAOS	Distributed Asynchronous Object Storage
Cray	Cray Inc.
ECI	Exascale Computing Initiative
ECP	Exascale Computing Project
FIT	Failure in Time
FOM	Figure of Merit
HBM	High-Bandwidth Memory
HI	Hardware and Integration
HPC	High-Performance Computing
HPE	Hewlett Packard Enterprise
HSU	Heterogeneous System Use Models
IBM	International Business Machines
Intel	Intel Corp.
IP	Intellectual Property
KPP	Key Performance Parameter
L3	Level 3
LLNS	Lawrence Livermore National Security
MIT	Memory Interface and Technologies
MPI	Message Passing Interface
NVIDIA	NVIDIA Corp.
NVMe	Non-Volatile Memory Express
OCC	Optimized CPU Chiplets
OSC	Open, Scalable Chip-to-Chip Interconnects
POC	Point of Contact
QoS	Quality of Service
R&D	Research and Development
RDMA	Remote Direct Memory Access
RFP	Request for Proposals
SDC	Silent Data Corruption
SME	Subject Matter Expert
SOW	Statement of Work
SVE	Scalable Vector Extension
TR	Technical Representative
USB	US Government
VCSEL	Vertical Cavity Surface Emission Laser
WBS	Work Breakdown Structure
WG	Working Group

EXECUTIVE SUMMARY

The PathForward element of the Exascale Computing Project (ECP) Hardware and Integration (HI) focus area has met its goals of preparing the US industry for exascale system procurements and generally improving US competitiveness in the worldwide computing market.

A competitive PathForward RFP (Request for Proposals) was released in 2016 seeking responses that would improve application performance and developer productivity while maximizing energy efficiency and reliability of an exascale system. Following a rigorous review process, six responses were selected for award and contract negotiations began. All six selected responses successfully led to contracts that were awarded and announced in June 2017. The six awardees were Advanced Micro Devices (AMD), Cray Inc. (Cray) (subsequently acquired by Hewlett Packard Enterprise [HPE]), HPE, International Business Machines (IBM), Intel Corp. (Intel) and NVIDIA Corp. (NVIDIA). All have successfully completed all contract milestones.

Each PathForward contract Statement of Work (SOW) details the work to be performed, milestones and deliverables for each milestone, and the planned start (authorization) and completion dates for the work to be completed for each milestone. The initial set of milestones were authorized upon award and all milestones were authorized for performance. Reviews of the PathForward work and deliverables were held bi-annually throughout the period of the contracts. Prior to the COVID-19 pandemic, reviews of the progress of the awardees were held at Lawrence Livermore National Laboratory. The final reviews of the awardees successful completion of their contract goals were held in August and September 2020. These were held virtually to ensure the health and safety of all participants. The reviews were well attended by the DOE national laboratories; the laboratory technical representatives responsible for the six PathForward projects; the ECP senior leadership team, including the ECP chief technology officer; and federal employees from DOE and other US agencies.

The quantity and breadth of the work and milestones across the six contracts presented a challenge for ECP to ensure the work was properly reviewed and feedback provided to the contract awardees in a timely fashion. The PathForward program engaged working groups for each contract, with members from each of the six core ECP DOE national laboratories tasked with ensuring the reviews were performed effectively by subject matter experts from across those institutions.

As of October 13, 2021, the 267 milestones across all projects are completed. The number of milestones completed by the awardees are not comparable to each other because the nature and scope of the milestones vary substantially. Thus, the number of milestones and their costs varies both within and, particularly, across the PathForward projects.

This report is a companion to the deliverable for the ECP milestone PM-HI-1040, Assess PathForward Impact Against Exascale Hardware Challenges, which gives a summary of the final status of each PathForward project, describes progress achieved against PathForward contract milestones, and includes a final assessment of each vendor's progress on key exascale challenges. This companion document details the results of the PathForward research to the extent possible without disclosing proprietary information and the impact on products and US exascale systems. In addition, it captures lessons learned in order to inform future projects in general and in high-performance computing in particular.

1. INTRODUCTION

The PathForward element of the Exascale Computing Project (ECP) Hardware and Integration (HI) focus area has met its goals of preparing the US industry for exascale system procurements and generally improving US competitiveness in the worldwide computing market. It has already achieved its key performance parameter (KPP) threshold and will likely meet its KPP objective by the time this report has been approved.

The ECP has accelerated the delivery of a capable exascale computing ecosystem. Reaching this level of computing has significant challenges partly due to the physical limits of existing computing hardware. To overcome the hardware limitations, the ECP's PathForward element invested in US companies for hardware research and development (R&D) with the objectives of preparing US industry for DOE exascale system procurements in the time frame of ECP and generally improving US competitiveness in the worldwide computing market. KPP-4 is designed to measure the ECP's success in accelerating US high-performance computing (HPC) companies ("vendor") innovation needed to deliver exascale computing systems. KPP-4 is defined as the number of milestones completed in the PathForward element relative to the total number of milestones contracted for completion by US HPC vendors (see below).

KPP ID	Description of Scope	Threshold KPP	Objective KPP	Verification Action/Evidence
KPP-4	Enrich the HPC hardware ecosystem	Vendors meet 80% of all the PathForward milestones	Vendors meet 100% of all the PathForward milestones	Independent review of the PathForward milestones to assure they meet the contract requirements; evidence is the final milestone deliverable

A competitive PathForward Request for Proposals (RFP) was released in 2016. Following a rigorous review process, six responses were selected for award and contract negotiations began. All six selected responses successfully led to contracts that were awarded and announced in June 2017. The six awardees were Advanced Micro Devices (AMD), Cray Inc. (Cray) (subsequently acquired by Hewlett Packard Enterprise [HPE]), HPE, International Business Machines (IBM), Intel Corp. (Intel) and NVIDIA Corp. (NVIDIA). The PathForward contracts were placed early in ECP's existence. This time frame was essential for meeting the broader goals of PathForward, which are discussed subsequently in this report. This time frame was enabled by the prior activities conducted as FastForward I and II and DesignForward I and II, which were earlier DOE Office of Science and National Nuclear Security Administration extreme-scale computing activities predating the DOE Exascale Computing Initiative (ECI).

Each PathForward contract statement of work (SOW) details the work that was to be performed, milestones and deliverables for each milestone, and the planned start (authorization) and completion dates for the work to be completed for each milestone. The initial set of milestones were authorized upon award. All of the milestones were authorized and all but two completed. Reviews of the PathForward work and deliverables were held bi-annually throughout the period of the contracts. Further, the contracts include intellectual property (IP) terms that allow the awardees to retain the rights to the IP that results from the work in exchange for significant cost-sharing investment (40% or more) by them.

All PathForward contracts were placed by Lawrence Livermore National Security (LLNS), LLC, the contractor for Lawrence Livermore National Laboratory (LLNL), as R&D contracts. These contracts use terms and conditions that recognize the long-lead nature of R&D activities, such as those conducted under

PathForward. As such, the contracts allow significant flexibility in adapting to lessons learned during the course of the work. Thus, milestone results often produce greater success than would otherwise be achieved and the results of the work are evaluated for their overall impact rather than strict conformance to quantitative goals. For example, as Section 1.3.2 discusses, projects set quantitative target goals but those targets were intentionally challenging and unlikely to be met universally.

This report is a companion to the deliverable for the ECP milestone PM-HI-1040, Assess PathForward Impact Against Exascale Hardware Challenges. The ECP milestone requires a final assessment of each PathForward project in the work breakdown structure (WBS) 2.4.1 portfolio. That deliverable report describes progress achieved against defined PathForward contract milestones and figure of merit (FOM) targets that provided guidance for the PathForward work to support broader ECP goals. The FOM targets were aggressive and as a result not all were achieved. Nonetheless, the work met the overall PathForward goals as planned. US exascale systems will benefit from the work and all the projects have impacted (and improved) the current product roadmaps of the awardees.

1.1 FINAL REVIEW

The final reviews of the progress of the awardees, including presentation of the overall work performed, were held virtually due to the COVID-19 pandemic during the weeks of August 10, 2020 and August 31, 2020. The reviews were well attended by the DOE national laboratories, as well as by federal government employees from DOE and from other agencies. Each awardee presented their work and its impact on their products and roadmaps, providing an opportunity for the attendees to ask detailed questions. Table 1 shows the schedule for the review.

Table 1. PathForward Q3 CY2020 Review Schedule.

Date	Awardee
Tuesday, August 11, 2020	NVIDIA
Wednesday, August 12, 2020	IBM
Thursday, August 13, 2020	Intel
Tuesday, September 1, 2020	AMD
Wednesday, September 2, 2020	HPE
Thursday, September 3, 2020	HPE (formerly Cray)

A special closed session was held after each vendor's session with the DOE and federal employee attendees to discuss the material presented. Notes taken during these sessions were used to develop the milestone report. Section 2 discusses the success of each PathForward contract and assesses the impact of the work.

1.2 ECP PATHFORWARD ROLES AND ACTIVITIES

ECP PathForward roles include the ECP WBS Level 3 (L3) control account managers (CAMs), technical representatives (TRs), subject matter experts (SMEs), and working groups (WGs). The PathForward WGs for each contract included members from each of the six core ECP DOE national laboratories tasked with ensuring that the reviews were performed effectively.

1.2.1 Technical Representatives (TRs)

PathForward TRs had primary responsibility to guide and to manage industry partner PathForward activities. Two TRs combined to perform the management functions for ECP; one person served as the ECP TR while an LLNL staff member served as the LLNS TR (except in one case in which the same individual performed both roles). Each ECP TR had primary responsibility for managing interactions with one of the awardees while each LLNS TR had final technical authority over all contractual matters for one awardee's activities since all PathForward contracts were placed through LLNL. With the L3 CAM's assistance, the TRs for each awardee negotiated the awardee's SOW and any required modifications to that SOW throughout the performance period of the contract.

1.2.2 Working Groups (WGs)

A WG focused on the activities of one awardee to assist the TRs for that awardee. Each WG consisted of a point of contact (POC) from each of the E6 laboratories (Argonne National Laboratory, Los Alamos National Laboratory, Lawrence Berkeley National Laboratory, LLNL, Oak Ridge National Laboratory, and Sandia National Laboratories) as well as a POC from the ECP Application Development focus area's Proxy Applications project. The ECP and LLNS TRs served as the POCs for their laboratories.

The WG met regularly with the awardee to assess progress and risks for that awardee's activities. It also served as the interface that ensured appropriate DOE SMEs were engaged in those activities. The POCs distributed the partner's milestone reports to SMEs at their laboratory and gathered feedback on the correctness, completeness, and merit of the work described in the milestone reports (as well as performing their own assessment of each report). The Proxy Application POC also assessed milestone reports. However, their primary role was to help ensure that appropriate DOE application SMEs were engaged in the awardee's activities, including assessing the milestone reports. The Proxy Application POC also guided the use of proxy applications (and real ones) by the awardee.

1.3 OVERALL EVALUATION OF THE SUCCESS OF PATHFORWARD

PathForward success can be evaluated along several axes that capture accomplishments of its overall goals. This report considers the degree to which PathForward has met its overall goals and the extent to which it will meet KPP-4, with which PathForward is associated. The report also assesses the extent to which the individual PathForward projects have contributed towards the overall success of PathForward.

PathForward has clearly been successful and, in general, each of the projects is considered to have been successful. With respect to KPP-4, the threshold value of acceptance of 80% of the total milestones of the projects was met early in FY20 and the objective value of 100% completion of those milestones was met in the first half of FY21. More importantly, the overall evaluation is that PathForward has met its goals to enable the first US exascale systems and to enhance the competitiveness of the US HPC industry. The rest of this section provides the current status of KPP-4 and a summary of how PathForward met those goals.

1.3.1 Overall Milestone Status

The work in each contract was defined by work packages. A work package represented a body of work related to a component of a system—such as compute cores, memory systems, interconnects, or nodes—or to a key system-level aspect for achieving ECP goals. Each work package consisted of multiple milestones to track and to assess progress of the work package. The TRs were responsible for managing assessment of milestone reports that document the work performed for each milestone.

After the awardee submitted a report, the milestone assessment process consisted of a detailed review by the TRs, the associated WG, and other SMEs with knowledge of the technology covered by the particular milestone. The review would identify any deficiencies in the report or the work that it documents with respect to the specific milestone acceptance criteria defined in the SOW. If any deficiencies were found, the TRs would provide feedback on them to awardees, who would perform any additional work required to address the deficiencies and then submit a revised report that documented the resolution, resulting in further review of the report and, potentially, additional feedback. Once any and all deficiencies were addressed, the TRs accepted the milestone.

Table 2 summarizes the overall progress of PathForward contract milestones as of January 25, 2021. Over 99% of the milestones have been completed (four are pending payment but counted as completed since they have been accepted), surpassing the KPP-4 threshold of 80%. As detailed in the individual sections, the last two have been submitted and will be completed soon, which would meet the KPP objective. Both are expected to be accepted by the end of February 2021.

Table 2. PathForward Milestone Status.

Milestones	Quantity
Total	267
Authorized	267
Accepted	267

1.3.2 Figures of Merit (FOMs)

In order to help ECP to determine the extent to which proposed activities would contribute to the achievement of overall ECP goals, responses to the PathForward RFP identified FOMs that would be achieved if the specific work packages were selected for funding. FOMs identified low-level goals related to the specific technology area addressed by the work package. These goals documented improvements beyond the planned products roadmaps in key metrics related to the technology area that would be targeted if the work package was funded. These target goals were aggressive, with the expectation that if all projects met all of the proposed FOMs then ECP would easily exceed its goals. Thus, overall ECP goals would be achieved even if some individual FOMs were not attained despite significant progress beyond product roadmaps that did not include PathForward funding.

The FOMs only serve as one mechanism for assessing the overall success of specific work packages. When appropriate and possible, each awardee provided final assessments of the progress towards the proposed FOMs as part of their final review. As indicated in Section 1.3.1, some milestones have not yet been completed, which impacted the ability to assess progress towards proposed FOMs during the associated final review. While the milestone report discusses progress towards FOMs when available, their absence has not impacted the assessment of the degree to which the individual projects have contributed to meeting overall ECP goals as discussed in the following section.

1.3.3 Assessment of Higher Level PathForward Goals

PathForward has had extensive positive impacts on the anticipated US exascale systems (Aurora at Argonne National Laboratory, Frontier at Oak Ridge National Laboratory and El Capitan at Lawrence Livermore National Laboratory) as well as on Perlmutter, the anticipated pre-exascale system at Lawrence Berkeley National Laboratory and Crossroads, the anticipated pre-exascale system at Los Alamos National Laboratory. While the selected systems do not include technologies from all of the awardees, PathForward research significantly increased the competitiveness of several RFP responses for those systems and resulted in significant benefits to the selected systems.

The technologies to be used in the five upcoming systems demonstrates the significant impact of PathForward funding on DOE systems. This impact is expected to increase as DOE national laboratories and possibly other US agencies acquire other systems after those five. All of the systems are expected to use the Cray Slingshot network that Cray (now part of HPE) developed in part through PathFoward funding. Frontier and El Capitan will include AMD processor technology (CPU and GPU) that incorporates several results of PathForward-funded research, much of which also contributed to the CPU technology planned to be used in Perlmutter. Aurora is anticipated to include Intel technology developed partly through PathFoward funding. Results of the PathForward funding of NVIDIA contributed to several improvements for GPUs that will be used in Perlmutter including in their reliability and in tools to analyze performance of applications on them.

Importantly, the technologies being deployed in the five upcoming systems that reflect PathForward funding are all standard product offerings and contribute to key aspects of the product roadmaps of the awardees. Additional influence on future products of the PathForward awardees is anticipated. For example, results of the research conducted by IBM under PathForward funding is expected to appear in their future CPU technology. Further, many of these products have been featured in recent announcements of system procurements not only in the US but also around the world, including in Europe and Australia. HPE/Cray has been particularly successful in the recent announcements and the systems regularly incorporate AMD processor technology to be used in the DOE systems. Overall, PathForward has clearly succeeded not only in substantially improving the systems that DOE will procure but has also substantially increased US competitiveness in high-performance computing. Thus, PathForward has achieved its high-level goals.

2. ASSESSMENT

This section assesses each PathForward contract to the extent possible without revealing proprietary information. The milestone report includes a summary of the authorized and completed milestones, a brief assessment of each work package, including notes from the final review, and any risks and challenges that arose and how they were addressed. Most importantly, the assessments include evaluations of the progress towards the target FOMs and the impact of the work on awardee products and roadmaps. In addition to the non-proprietary companion to the milestone report, each of the awardees has provided a general, non-proprietary statement that details the activities performed under PathForward.

2.1 AMD

2.1.1 Description

AMD had six work packages. DOE investments have enabled AMD to provide immense computational power in America's pre-exascale and exascale supercomputers through innovative computing solutions, acceleration of key hardware and software technologies, the rejuvenation of the HPC hardware and software ecosystem, and joint research and co-design with the DOE national laboratories. Below is a short description regarding each work package.

2.1.1.1 Compute-Optimized GPUs (COG) Work Package

The Compute-Optimized GPUs (COG) work package developed new GPU capabilities and enhancements to better optimize AMD GPU performance for HPC workloads based on co-design utilizing the ECP proxy applications. It developed software optimizations for DOE proxy applications to execute more efficiently on commercially available AMD GPUs to increase performance. Finally, it explored and

innovated new GPU implementation techniques such as circuits and power management mechanisms to significantly reduce GPU power consumption.

2.1.1.2 Optimized CPU Chiplets (OCC) Work Package

The Optimized CPU Chiplets (OCC) work package enabled CPU optimizations tuned for important HPC workloads through detailed characterization of DOE proxy applications. It developed improvements in performance, energy efficiency, and data movement in the CPU cache and memory hierarchy. Finally, it explored new performance-enhancement mechanisms in key CPU micro-architecture structures supporting speculative and superscalar instruction execution.

2.1.1.3 Component and Memory Integration (CMI) Work Package

The In-Package Component and Memory Integration (CMI) work package conducted studies of the architectural integration of heterogeneous computing components such as CPUs and GPUs that have directly motivated and influenced the design of AMD's exascale node architecture. Its evaluation of physical integration technologies provided key findings and guidance for appropriate and effective packaging for AMD node designs with Accelerated Processing Platform (APP) components.

2.1.1.4 Open, Scalable Chip-to-Chip Interconnects (OSC) Work Package

The Open, Scalable Chip-to-Chip Interconnects (OSC) work package employed early prototyping efforts to provide key insights and experiences in the software implications of advanced interconnect technologies, and into the challenges and solutions related to interfaces between commercially available AMD hardware and external interconnect standards and protocols. Importantly, this work package enhanced a collaborative relationship with a key industry partner for interconnects.

2.1.1.5 Heterogeneous System Use Models (HSU) Work Package

The Heterogeneous System Use Model (HSU) work package conducted studies of the coherence, communication, and data movement on exascale nodes composed of discrete compute components that have impacted the design of AMD's exascale node architecture. Its co-design between AMD, system integrator partners, and the DOE have driven a design-space reduction and down-selection for practical and effective exascale node and system architectures. Further, its analysis of network integration has led to optimized GPU-network communication designs.

2.1.1.6 Memory Interface and Technologies (MIT) Work Package

The Memory Interface and Technologies (MIT) work package detailed memory roadmaps and requirements for exascale systems to guide the selection of viable memory options for inclusion in exascale systems. Its examination and analysis of exascale requirements have driven and influenced memory vendor roadmaps and plans.

2.1.2 Overall Assessment, Including FOM Progress and Product Impact

DOE reviewers have consistently given AMD high marks for their DOE-funded hardware R&D projects, including FastForward, DesignForward, and PathForward. In general, their work packages have been well integrated and presented a consistent vision of exascale computing. Their PathForward work has influenced product designs that led to their successful participation in Collaboration of Oak Ridge, Argonne and Livermore (CORAL)-2 proposals. It has also significantly influenced and enhanced AMD roadmaps and products beyond those proposals. The project is completed and has no outstanding risks.

2.2 HPE (FORMERLY CRAY, INC.)

2.2.1 Description

HPE (formerly Cray) had four work packages. The most critical goals of the project were:

- Increasing the power efficiency of CPU-based architectures, both through core-level design optimization, as well as efficiently capturing performance across the node;
- Driving down the cost and technical risk of deploying advanced memory technologies, including high-bandwidth memory (HBM);
- Developing a network that can provide advanced message passing interface (MPI) performance at the massive scale required for exascale systems; and
- Enhancing that network to address the requirements of large-scale converged HPC/AI/Analytics workflows running across a single system, including new levels of job isolation and quality of service guarantees.

The following sections provide high-level overviews of Cray's work packages.

2.2.1.1 Next-Generation Network Development and Optimization

Through PathForward, along with funding from other programs, Cray developed Slingshot, an HPC-enabled Ethernet-compliant network, with an emphasis on application and workload needs of exascale science. Specific features were developed to provide enhanced performance and scalability of workloads, including aggressive MPI and remote direct memory access (RDMA) capabilities, as well as enhanced Quality of Service (QoS). The Slingshot network will provide the backbone of future HPC and AI products from HPE/Cray, where it will have a broad impact on HPC and datacenter network capabilities. Slingshot will also deliver the scalable performance needed to make DOE's upcoming systems successful.

2.2.1.2 Future Network Architecture

PathForward also built on the development of Slingshot-1 to develop the architecture for an open, standards compliant, second-generation Slingshot network capable of further improving performance for exascale systems. Of particular focus was the advancement of features enabling emerging AI workloads, as well as workflows combining simulation, AI, and data analytics. This work will directly impact the features and functionality of future Slingshot networks developed by HPE.

2.2.1.3 High-Performance Next-Generation Instruction-Set Architecture and CPUs

The Scalable Vector Extension (SVE) was introduced in the Arm architecture to improve performance and energy efficiency in HPC and machine learning. At the time of definition of the PathForward project, several improvements to SVE were in-flight candidates to improve performance for a larger set of workloads and use cases, which led to the release of SVE2. Some of these features for HPC were evaluated in the context of FastForward-2. The PathForward program funded research on several proposals of architecture extensions that are now part of a technology cache for future Arm ISA enhancements. Features specific to vector processing, such as richer strides, segmented scans, and dense matrix/AI acceleration showed important gains for ECP workloads, as did transactional memory and cache stashing. These analyses feed into candidates for future architecture extensions to enable the Arm ecosystem to improve performance and energy efficiency for a broader range of exascale workloads.

HPE (formerly Cray) also worked closely with Marvell on this work package. Many DOE applications feature sparse computations that are implemented through gather and scatter memory accesses when vectorized. Gather/scatter support in CPUs has traditionally been weaker than contiguous accesses for a variety of reasons related to the implementation of load/store units and memory subsystems in modern CPUs. Using PathForward investments, Marvell investigated new approaches for gather/scatter that target actual use cases in DOE codes to optimize the number of operations and time taken to process these instructions. These optimizations will be implemented in future ThunderX processors and will contribute to more efficient execution of a significant set of codes that feature gather/scatter constructs extensively.

2.2.1.4 Development of Cost-efficient Multi-die Packaging Technologies

With the abatement of Moore's law gains, alternate approaches are required to maintain the rate of increase of computational capacity in a package to reach exascale and beyond. One approach that has shown promise is packaging multiple dies together on a common substrate or interposer. Such packaging also allows colocation of memory stacks with compute dies to enable much higher memory bandwidths than are critical to HPC science performance. Using PathForward investments, Marvell and Cray investigated a range of new packaging technologies and memories, with the knowledge gained from these studies expected to enable innovative combinations of compute and memory. PathForward also drove investigations into how memory configurations within each socket impact system-level metrics such as performance, power, and reliability.

2.2.2 Overall Assessment, Including FOM Progress and Product Impact

Cray has consistently been a strong performer with previous DOE funding programs and this strong commitment continued under PathForward. The team was well resourced and was populated by staff with considerable product and research experience. Unlike a number of vendors, Cray's product groups are incorporated into the research teams from the outset and this pays dividends through strong requirements analysis and reduced time to market. The HPE acquisition of Cray in September of 2019 did not negatively impact the deliverable timeline or the quality of the work. The project is completed and has no outstanding risks.

2.3 HPE

2.3.1 Description

HPE had five work packages, which address these technology areas: system architecture, node design, data movement, optical interconnects and I/O node design.

2.3.1.1 System Architecture Work Package

HPE is an industry leader in large-system designs. Moving to exascale provides significant challenges to existing system architectures. HPE sought to develop a more balanced system architecture surrounding memory semantic fabrics and restructured proxy applications to better utilize these new fabric architectures. HPE also developed distributed monitoring solutions that scale to exascale to replace more centralized approaches. Finally, HPE explored the use of lightweight, thermally conductive plastics to enable more economical and efficient cooling designs and to support a fan-less liquid cooling solution.

2.3.1.2 Node Design Work Package

HPE's node research extended load/store memory semantics from the processor-memory interface to the system fabric. HPE developed the first performant memory semantics chipset to implement the Gen-Z

protocol. HPE pushed the boundaries of silicon module design creating one of the most complex set of multi-chip modules in existence. The modules combine high-speed digital circuits and optical interfaces in the same package. HPE worked with volume silicon foundries to ensure that the technology would be economically viable going forward.

2.3.1.3 Data Movement Work Package

An exascale system presents challenges to existing fabric topologies. HPE developed modeling tools to run networking simulations of exascale interconnects more accurately and efficiently. HPE investigated new fabric topologies such as Hyper-X with new congestion management algorithms and showed that their performance improves upon existing topologies. HPE designed and built a high-radix router to evaluate the accuracy of their models and to improve performance prediction. HPE also developed extensions to the Gen-Z protocol to provide better message passing efficiency as well as other scalability improvements critical to high-performance computing.

2.3.1.4 Optical Interconnects Work Package

For power and performance viability, exascale systems will increasingly depend on integrated photonics. HPE pursued the integration of two types of silicon photonics into CMOS: silicon photonics based on micro-ring oscillators; and photonics based on vertical cavity surface emission lasers (VCSELs). HPE worked through volume silicon foundries to enable the technology for potential volume applications. The VCSEL technology was used in the fabric bridge and switch chipset modules, while the silicon photonics technology was tested in standalone prototypes. HPE also developed on-chip packaging technologies that enable direct connection between optical fibers and silicon chips.

2.3.1.5 I/O Node Design Work Package

Requirements for additional I/O bandwidth in exascale computing—for example due to increased checkpointing needs—necessitate a new approach to the design of the I/O subsystem. HPE addressed this need through a hybrid media controller that provides a combination of fabric attached load/store memory and SSD-backed storage. The memory controller includes a memory side accelerator to offload some of the data movement and storage functions to the I/O module itself. HPE also developed a set of prototype file system services that capitalize on the many features of the I/O node design capabilities and implement novel features such as transparent tiering and multi-node resilience.

2.3.2 Overall Assessment, Including FOM Progress and Product Impact

PathForward was one of the largest R&D programs upon which HPE has embarked during the last decade. HPE emphasized the significant value that the PathForward program provided, in terms both of technology advancement and of increased understanding of DOE applications and operational practices through workshops and deep dives. Overall, the PathForward project has been a catalyst to jump-start HPE's technology toward post-exascale computing and the work packages are expected to impact future HPE system architectures and networking solutions. HPE has stated that the PathForward program and the co-design collaboration model has provided HPE and its partners the focus and direction needed to address the exascale computational requirements while also enhancing US leadership and competitiveness in the global IT industry. The project is completed and has no outstanding risks.

2.4 INTEL CORPORATION

2.4.1 Description

Intel had four work packages. Each work package is discussed in further detail below.

2.4.1.1 High-Performance and Energy Efficient Computing

Key requirements to reach exascale system goals are improvements in performance and performance per watt per node. Intel investigated novel hardware compute architectures combined with a fully prototyped software tool-chain. A key tenet of this work package was to use existing and widely applicable programming models. Collaboration with DOE, using ECP applications, at codesign workshops and other avenues led to an architecture with demonstrable advantages and helped Intel develop a highly energy-efficient and disruptive computing paradigm with potential impact to HPC.

2.4.1.2 High-Performance Communication

Scalability of parallel performance is a growing concern as we continue to increase the number of cores on a die or in multi-socket coherent nodes as on-node core counts continue to increase. HPC applications are particularly sensitive to this phenomenon due to their parallel nature and high bandwidth demands on memory and cache coherence systems. PathForward funding has enabled exploration and evaluation of key technologies that improve communication and memory sharing among threads, ranks, processes within nodes, as well as core-to-device interactions for distributed memory systems. Through a combination of execution-driven simulation, and the use of DOE benchmark codes, this work package has supported potential adoption of these features into future generation roadmap products.

2.4.1.3 High-Speed and Scalable Interconnects

Performance of exascale HPC systems depends critically on the interconnect delivering high bandwidth and low latency, while minimizing the total available system power. Traditionally, fabric solutions are built with edge connected cables and transceivers, but this approach is becoming inefficient in terms of power, as the overhead of long signal routing from switch IC to front-plate consumes a larger and larger share of power as bandwidth increases. Intel has addressed the power efficiency challenge by developing co-packaged optical interconnect technology that integrates high-bandwidth photonic engines with next-generation high-bandwidth switches. This optical co-packaging approach addresses the bandwidth density challenge through high-channel count scaling of silicon photonics optical components, the bandwidth and power consumption challenges through lower loss electrical channels, and the cost challenge through integration. The PathForward project helped fund an intensive technology development effort that resulted in the demonstration of a fully functional Ethernet switch with co-packaged optics, demonstrating 400 Gbps optical I/O links interoperable with a commercial switching system.

2.4.1.4 Scalable I/O

High fidelity applications on exascale systems are expected to generate data at extreme speeds with high resolution. The Distributed Asynchronous Object Storage (DAOS) software provides scientific applications with low latency, high bandwidth, and fine-grained access to system-wide storage. The PathForward program funded the development of a number of core DAOS features that will be deployed on the Aurora system. This includes Non-Volatile Memory Express (NVMe) SSD support, scalable service monitoring and a more efficient way to wire-up the DAOS service. Moreover, PathForward funding allowed exploration of a new model that integrates storage nodes into the compute rack without sacrificing storage resilience or reducing compute resources. A new storage acceleration framework based

on FPGAs has been prototyped and can be leveraged to provide faster and more capable (i.e., computational storage) storage systems.

2.4.2 Overall Assessment, Including FOM Progress and Product Impact

PathForward has helped enable Intel to develop disruptive technologies that will deliver exceptional energy efficiency and cost performance. These technologies include improvements in Intel's roadmap capabilities for compute and communication. Some of the work will impact Argonne's exascale system. All of the work is expected to impact subsequent systems. Intel actively engaged application teams to ensure a broad evaluation of their technologies and that led to important feedback that Intel incorporated into its development work and planning. The project is completed and has no outstanding risks.

2.5 IBM

2.5.1 Description

IBM's PathForward project comprised ten work packages. IBM developed and validated a configurable, flexible and efficient exascale solutions architecture, suitable for the large spectrum of complex workflows that require exascale capabilities, and which enabled further advances in scientific and enterprise computing, both on-premises and in the cloud. The work packages are discussed in further detail below.

2.5.1.1 Resilient Heterogeneous Compute Framework

The Heterogeneous Frameworks work package investigated approaches to building systems that allow reconfigurability of system resources at job launch, increasing overall system utilization and resilience.

2.5.1.2 Extended Memory

The Extended Memory work package focused on enhancing programmability and usability by providing symmetric access to all memories available in a reconfigurable system. This package studied the challenges involved in supporting system-wide unified addressing and scalable coherence and consistency, in hardware as well as in system and application software.

2.5.1.3 Active Network

The Active Network topic comprised two work packages. The first addressed the development of improved messaging engines to offload data transfer tasks from the CPU. The second work package investigated further development of InfiniBand technology to support the reconfigurable compute framework and its unified addressing, coherence and consistency requirements.

2.5.1.4 Open Coherent Interface

The Open Coherent Interface work package evaluated open interface standards and improvements that can be adopted by the industry, so that a spectrum of solutions will be available for inclusion in exascale and post-exascale systems, supporting very tight coupling and co-design of compute elements (both processor and accelerators), memory, storage class memory, and networking.

2.5.1.5 Processor Core Architecture

The Processor Core Architecture work package studied and validated fundamental improvements in processor design to allow inclusion in exascale and post-exascale processors.

2.5.1.6 Memory Interfaces and Technologies

The Memory Interfaces and Technologies work package developed new memory interfaces that support increased bandwidth at reduced pin count. The package targeted fundamentally new interfaces that IBM seeks to standardize.

2.5.1.7 Electrical and Optical Interconnects

The Electrical and Optical Interconnects were the subjects of two work packages. They addressed power-efficient high-bandwidth electrical signaling improvements and proof-of-concept optical networking and signaling using high bandwidth, low latency interfaces.

2.5.1.8 Innovative Power Delivery Systems

The Innovative Power Delivery Systems work package focused on improving power distribution efficiency and packaging and thereby improving the TCO of future computing systems.

2.5.2 Overall Assessment, Including FOM Progress and Product Impact

PathForward influenced and effectively sponsored new and innovative technologies that will benefit not only DOE applications and workloads, but also commercial customer applications in the 2021–2028 timeframe. In particular, prototypes generated in the different work packages are being evaluated for adoption in future generations of microprocessors and systems, including hierarchical frameworks, extended memory, active messaging engines, open interfaces, compute acceleration engines, memory buffers, optical interconnects, and power conversion. The project is completed and has no outstanding risks.

2.6 NVIDIA

2.6.1 Description

NVIDIA's PathForward work was divided into two work packages. One work package was focused on resilience. The other was focused on performance improvements for energy efficiency and performance analysis.

2.6.1.1 Resilient Exascale Computing

One of the fundamental challenges in scaling up current technologies to exascale is reliability. When the NVIDIA PathForward project started, the number of GPU compute nodes in an exascale system was estimated to be many tens of thousands. At that time, NVIDIA estimated that silent data corruption (SDC) failures in time (FITs) would be sufficiently high that the resulting mean-time-to-failure would be inadequate to meet exascale system requirements. The PathForward program funded the development and implementation of a number of enhancements to NVIDIA’s methodology and hardware that are expected to reduce the SDC FIT rate and to improve availability by a large factor (and to meet exascale system requirements) over NVIDIA’s current generation of products, while only incurring a small silicon area overhead.

2.6.1.2 Energy Efficient Exascale GPUs and Next Generation GPU Application Analysis

To achieve the performance and energy targets of exascale systems, GPU efficiency must dramatically increase. While current GPUs are extremely energy efficient, achieving exascale power goals requires chip level efficiency that is significantly higher. NVIDIA's approach to improving energy efficiency was to characterize the power and performance of a range of GPU accelerated technical computing applications on existing systems and to identify a range of improvement opportunities throughout both the GPU memory system and processing elements. Ultimately, the PathForward program funded the development and implementation of several key new features in future NVIDIA GPU architectures.

To achieve the energy-efficiency and resiliency gains required to enable exascale applications, NVIDIA additionally identified the need to accelerate progress by bolstering the data-driven process that fuels NVIDIA's hardware and software development. Insights into detailed application characteristics and behaviors in GPU systems guide the conception of new features and decision making at every level of engineering from circuit design to application software. The ability to extract and analyze data from GPU applications presents two key challenges: data collection techniques that accurately capture interactions between distinct devices and tasks running asynchronously in modern, tightly-integrated HPC nodes; and specialized analysis use-cases that existing tools could address. With PathForward funding, NVIDIA developed hardware and software capabilities that provide a foundation for robust solutions to these challenges and integrated them into application analysis tools. The results of this work continue to have a growing impact not only on the design of GPUs across NVIDIA's entire product portfolio, but also on their external tools that assist with optimizing systems and code.

2.6.2 Overall Assessment, Including FOM Progress and Product Impact

NVIDIA responded promptly and fully to all comments on their milestone reports. NVIDIA ensured that all key ECP participants had access to their results, which has helped to ensure that the work realizes its full potential value. NVIDIA also showed that they are willing to cooperate on questions not covered by their SOW: some ECP application developers raised questions about GPU suitability for certain types of applications, such as adaptive mesh refinement codes. NVIDIA agreed to investigate these concerns despite them being outside the scope of their SOW.

PathForward funded the accelerated development and implementation of several key new features in NVIDIA's future GPU architectures, allowing them to reach the enhanced resiliency, architectural efficiency, and analysis goals, well ahead of the natural progression of their development roadmap. The results continue to have a growing impact not only on the design of GPUs across NVIDIA's entire product portfolio, but also on the tools that it provides externally to assist with optimizing systems and code. The project is completed and has no outstanding risks.

3. RESOURCE REQUIREMENTS

The primary resource requirements for PathForward activities were the milestone funding. ECP resources involved in the PathForward activities also included: ECP TRs (one per contract at 25% effort for 1.5 FTEs total); LLNS TRs (one per contract at 15% effort, with one included in the previously stated ECP TR effort for 0.75 FTEs total); the additional Laboratory POCs in each WG (roughly 10% effort for each of 5 POCs at 5 laboratories for 2.5 FTEs total); and funding for SME engagement and milestone reviews (approximately 3.0 FTEs total). No DOE HPC resources were required.

4. CONCLUSIONS AND LESSONS LEARNED

Overall, the PathForward element achieved its goals. The awardees completed all milestones, thus meeting the PathForward KPP-4 objective. More importantly, PathForward research contributed several technologies that will be employed in the first US exascale systems and that, as a result, have significantly strengthened the US economic competitiveness and provided significant advantages to the US computer industry, here and abroad. Further, all awardees engaged with other ECP activities, which will help ensure that ECP meets its other KPPs.

The remainder of this section provides lessons learned from PathForward. At the highest level, two critical lessons emerge. First, sustained investment, as evinced in not only PathForward but also in FastForward, DesignForward, and non-recurring engineering contracts associated with system procurements is essential. The second critical lesson is the fact that sustained investment from the US government ensures the high likelihood that potential partners in critical industries will pursue directions that reflect the requirements of the US government. Industry partners who are succeeding without the investment are spurred to listen to DOE and other US government entities and to manufacture products that suit their workloads. Further, the success of AMD in the CORAL-2 procurements demonstrates that these investments enable increased competition that serve the interests of those US government entities by ensuring the highest quality offerings at the best possible price.

Overall, the awardees very much appreciated the opportunity to perform innovative research and development under PathForward funding. Much of this work comprised topics that they wanted to investigate in depth and that they believed would benefit customers such as DOE. PathForward funding helped justify the time and cost required for a proper evaluation to their corporate management. All awardees stated that they would like to find ways to continue work on similar topics and, in particular, to build on joint interagency projects.

Viewed through the lens of hindsight, some lessons are hard to put into practice. Awardees who picked the right technology directions are naturally viewed as the most successful. Other awardees picked alternative paths that may eventually prove valuable but have not yet had as much impact. Still others followed simpler, less innovative paths that impacted exascale systems but may have been produced without PathForward investment. These awardees may have found that their corporate culture would not provide room for failure, which is necessary for major innovations. Nonetheless, PathForward funded an appropriate overall mix of projects and technologies that range from extremely innovative and risky to straightforward and likely to provide useful (if relatively small) improvements.

At lower levels, the success of PathForward provides many lessons as well. On the positive side, the SOWs were well written and specified the work that was required without ambiguity. Milestone acceptance criteria provided clear statements on which to base acceptance. This clarity simplified interactions with vendors for PathForward TRs and other ECP participants. The awardees also appreciated having predictability in contract structure and execution. Nonetheless, flexibility in the DOE contracting process that allowed milestones to be restructured or even redirected was also considered essential. For projects that involved subcontractors or even cross-project interactions, the legal effort required to make sure that everyone was protected, from an intellectual property perspective, was difficult but worthwhile. Overall, the awardees generally commended DOE's contracting processes.

On the less positive side, ensuring engagement with the full range of projects was often difficult. SMEs did not always respond promptly to requests for milestone report feedback. In some instances, awardees found that milestone review cycles were too long. Awardees, as well as the TRs, agreed that richer, more consistent feedback on deliverables would have helped the projects be even more successful. SMEs were

often pressed with demands and deadlines from their other projects and could not respond promptly to sporadic reviewing tasks. Ideally, the laboratories would allocate more resources towards SME reviews, perhaps even funding multiple SMEs per laboratory to engage with each project for its entire duration. On a related issue, some projects asserted that they would have preferred that application experts thought more deeply about how to exploit proposed hardware features.

The use of mini-apps (e.g., proxy applications) proved to be a dual-edged sword. While their use clearly helped focus activities on DOE workloads, several awardees found that no appropriate mini-app existed to support explorations of hardware directions, such as network performance and low-level technologies, silicon photonics, or circuit-level power optimizations. Nonetheless, some awardees commented that they were presented with too many mini-apps for strong co-design and would have been better served if DOE aggressively filtered the list to choose one or two that could have been used over the entire PathForward project. They believed that a deep engagement with the application experts for the small set would have resulted in experiments with algorithmic changes as well as detailed information about the workloads. Further, proxy application identification should start earlier, even during the RFP preparation process.

Generally, awardees appreciated the close interactions that PathForward fostered. Some commented that work under PathForward was less independent than under FastForward and DesignForward. The in-person reviews ensured awardees regularly summarized their progress and results, as well as the impact of the work on their evolving roadmaps, enabling individual ECP participants to follow multiple projects closely. Nonetheless, awardees could have benefited from closer interaction with and more guidance from the ECP Application Development focus area, especially developers of ECP applications and proxy applications. Not all awardees successfully evaluated all their research products in the context of codes that are relevant to ECP. In some cases, they considered ECP applications, but late in the project. As stated above, the issue is partly due to a mismatch between the types of mini-apps that ECP provided (large, complex applications) and the types that were needed to inform some directions that the awardees were pursuing. Future activities such as PathForward should identify a range of mini-apps, including (in addition to the type used within PathForward): small, individual loops for hardware-simulation studies; full, multi-application workflows for system-wide design decisions; and simple, flexible expressions of applications for evaluating the potential of new programming models.

Despite these complications, the awardees found that ECP provided a common place to make applications and software projects visible and broadened the set of workload requirements that they considered. Generally, engagement with DOE application and software experts provided an important bigger picture for the work of the awardees while engaging DOE developers in pushing how technologies are used and exposed important aspects, both positive and negative, of their research directions.

4.1 RECOMMENDATIONS

The experiences of PathForward provide clear guidance for future R&D funding of the large-scale computing industry. The benefits demonstrate that the US Government (USG) can still exert significant influence on industry directions so that USG requirements for large-scale computing can be met. From these lessons we immediately identify that continued funding is critical and must involve tight collaboration of DOE national laboratory personnel with industry researchers.

The general structure of PathForward proved effective, although two changes might enable even greater impact. First, while a focus on hardware research is essential, DOE and industry participants consistently agreed that more direct explorations of system software implementations should also be funded in order to ensure realization of the benefit of the hardware research. In addition, while these projects followed the essential plan of funding all major US corporations involved in large-scale computing, the recent emergence of several start-up companies that are exploring innovative hardware directions, particularly in

relation to artificial intelligence (AI), indicates the USG would also be well served by funding some of these smaller companies. Thus, overall future programs should have expanded scope both in terms of the extent to which software is explored and the range and number of awardees.

A wide range of potential topics for future research also emerged from the projects. Besides the need to explore software directions more thoroughly, perhaps the most essential directions are to ensure that the USG reaps the benefits of AI-focused hardware in its large-scale systems and of the diverse range of devices already present in large-scale systems. These complementary directions would explore heterogeneous system architectures that include a range of compute node types, rather than the current trend towards systems limited to node-level heterogeneity. Such systems require continued advancements in processor (e.g., dataflow oriented processors commonly being pursued for AI acceleration), memory (volatile and non-volatile, both integrated with processors to ensure low latency and disaggregated to ensure high utilization) and networking (e.g., more tightly integrated silicon photonics) technologies.

5. ACKNOWLEDGMENTS

This research was supported by the Exascale Computing Project (ECP), Project Number: 17-SC-20-SC, a collaborative effort of two DOE organizations—the Office of Science and the National Nuclear Security Administration—responsible for the planning and preparation of a capable exascale ecosystem—including software, applications, hardware, advanced system engineering, and early testbed platforms—to support the nation's exascale computing imperative.