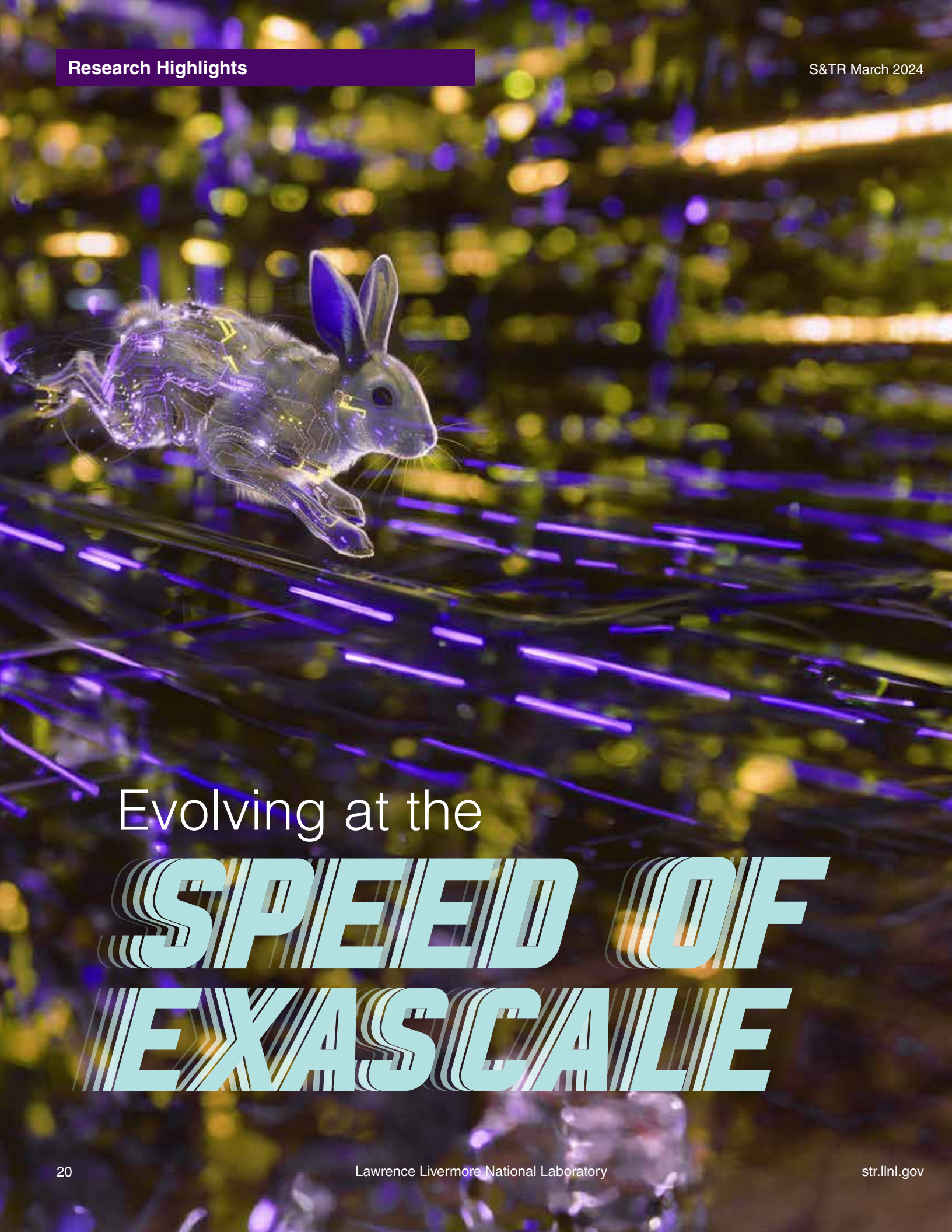
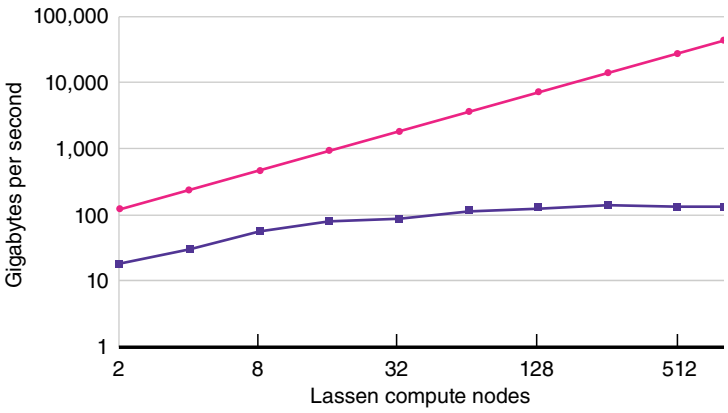


Evolving at the SPEED OF EXASCALE

HIGH-PERFORMANCE computing (HPC) speeds the design of new technologies, yet the path from start to finish is rarely smooth. Bugs, broken codes, or system failures require added time for troubleshooting and increase the risk of data loss. Lawrence Livermore, one of the Department of Energy (DOE) national laboratories at the forefront of HPC, has addressed failure recovery by developing the Scalable Checkpoint/Restart (SCR) framework.

SCR, developed in 2008, stores data mid-execution via checkpointing, a technique that enables streamlined data caching and eliminates the need for time-consuming manual failure recovery. SCR automatically restarts codes after a failure and does so in mere minutes. By managing the data on complex storage hierarchies, SCR handles both application data and checkpoints simultaneously and provides full-featured checkpoint/restart support. SCR features highly portable, production-level software that is easy to integrate into HPC applications. Livermore computer scientists earned a 2019 R&D 100 award for an iteration of the SCR framework that enables HPC applications to use hierarchical storage systems effectively. (See *S&TR,* July 2020, pp. 8–9.)

This crucial capability for exascale computing has become increasingly relevant as DOE's exascale systems—Frontier (Oak Ridge National Laboratory), Aurora (Argonne National Laboratory), and El Capitan (Lawrence Livermore)—come online. Livermore's SCR team has adapted the framework for exascale computing needs through collaboration with Oak Ridge and Argonne as well as through development of unique compute nodes called Rabbits that were designed with SCR's key capabilities in mind.



The Scalable Checkpoint/Restart (SCR) framework's input/output (I/O) techniques scale linearly with the number of compute nodes used to run an application, as shown for Livermore's Lassen supercomputer. The parallel file system (purple) reaches its scaling limit at about 64 compute nodes. SCR's I/O mechanisms (pink) are up to 1,000x faster than the parallel file system alone.

El Capitan, Lawrence Livermore's first exascale supercomputer, will come online in 2024 and is capable of performing more than two exaflops, or two quintillion double precision floating-point operations per second. Compute nodes are outfitted with specialized Rabbit nodes that will enhance the performance of SCR on the new computing system.

Speeding Simulations

Working toward ensuring SCR's application to future exascale computing needs at the Laboratory, Livermore's SCR team has adapted the framework and advanced its capabilities since its launch, keeping pace with the evolution of Laboratory computing systems from the 44.2-teraflop (trillion floating-point operations per second), cluster-based Atlas supercomputer to the 125-petaflop (quadrillion floating-point operations per second), heterogeneous Sierra supercomputer. In addition to supporting mission applications on Livermore's flagship supercomputing systems, scientists at DOE national laboratories such as Oak Ridge and Argonne have used SCR on their own applications, including on Frontier, the nation's first exascale computer. SCR has significantly improved the performance of input/output (I/O) operations—the reading or writing of data from a file system—to reduce the time to solution for large-scale simulations run on HPC systems. SCR uses fast storage tiers attached to the physical architecture of compute nodes on HPC systems, referred to as node-local storage, to cache application and resilience data quickly, speed operations, and minimize the risk of system failure—all of which makes running simulations on systems with higher failure rates possible.

Livermore computer scientist Ramesh Pankajakshan has run SW4, a code for simulating seismic wave propagation in three dimensions as a component of the EQSIM application, on Frontier. SW4 mimics seismic activity in a region based on what is known about the area's geological characteristics and history of earthquakes. The simulations are then compared to real-world data for accuracy and inform engineers designing new buildings or infrastructure. SCR plays a key role in these simulations.

Rabbits are specialized nodes that feature a processor, two peripheral component interconnect express switches, and two Slingshot Network Interface Cards that provide connectivity beyond the Rabbit to high-speed I/O components. Rabbit nodes are unique to El Capitan and were chosen by the Livermore high-performance computing architecture and planning team to help overcome challenges related to I/O at the exascale level.

"Without SCR making checkpointing more efficient, we can't finish the solution," says Pankajakshan. "If checkpointing takes 20 minutes per hour without SCR, we have 33 percent less computation time. SCR makes more computation possible. Idle time is minimized, and we, as users, are good stewards of the hardware and the taxpayers' dollars."

Adaptations for Exascale

The Laboratory's SCR team will support El Capitan's operations soon after the system comes online. To better understand how SCR will function on exascale computers, Livermore staff, including SCR team members, have applied their supercomputing expertise in partnership with colleagues at Oak Ridge to help launch Frontier and have more recently been collaborating with colleagues at Argonne to launch Aurora. Working on Frontier enabled Lawrence Livermore computer scientists to better understand the challenges related to transitioning from petascale to exascale systems, in particular failures and performance of I/O operations, and how those might apply to the 2024 launch of El Capitan. "We wanted to be proactive in talking to colleagues at Oak Ridge about what they were seeing so we could help. We can determine a lot about what we can do for the system by seeing it in its early stages of deployment," says SCR project founder Adam Moody.

El Capitan, which is expected to be the world's most powerful supercomputer when it comes online, will be capable of performing more than 2 exaflops, or about 2 quintillion double precision floating-point operations per second, making it roughly 16 times faster than its predecessor, Sierra. This faster speed renders SCR all the more important—the faster and more powerful the supercomputer, the greater the consequences if failures occur. "As supercomputers become bigger and bigger, the probability that anything could fail at any time is greater. SCR saves HPC users time and energy by managing many of the manual monitoring and restart tasks required when a failure occurs," says Kathryn Mohror, SCR project lead.

By employing storage tiers to cache applications and resilience data, SCR enables results in less time with lower consequence of failure, which means a lower risk of losing calculation potential. Preparing for the transition to exaflop computing has been one of the SCR team's biggest priorities since winning the 2019 R&D 100 award. One of the most remarkable innovations for the arrival of exascale computing at Livermore is the design and implementation of Rabbits.

Designed with El Capitan in mind, Rabbits are specialized nodes for the I/O operations that further improve SCR's capabilities. Each Rabbit node features a processor as well as peripheral component interconnect express (PCIe) switches and Slingshot Network Interface Cards (NICs), devices that connect the Rabbit node to other high-speed I/O components. In addition, users can customize how storage on Rabbit nodes is presented to applications via software commands. By specifying the best Rabbit configuration for application needs, users can gain better

The SCR Framework 2.0 team, including members from Livermore and other institutions (as noted), received their R&D 100 award at the 2019 ceremony: (from left) Kento Sato (RIKEN), Cameron Stanavige, Kathleen Shoga, Bogden Nicolae (Argonne National Laboratory), Kathryn Mohror, Elsa Gonsiorowski, Adam Moody, Greg Becker, and Greg Kosinovsky.

performance and take advantage of tools such as SCR. According to Bronis de Supinski, Livermore Computing's chief technology officer and El Capitan's architect, the Rabbits' ability to run user tools, including SCR, made a compelling argument for their inclusion in the new supercomputer. He says, "SCR can do part of its magic with shared storage, but SCR can have a broader reach when local and shared storage are configured on Rabbit modules. That capability stood out as something we wanted in El Capitan."

SCR achieves its file protection and management capabilities by creating copies of the files in the storage hierarchy, which protects them from data loss due to catastrophic failures, and transfers the files to the parallel file system for access after the job is finished. Marty McFadden, a computer scientist at the Laboratory, has worked to make SCR functional on the new Rabbit nodes. Two key SCR features available on Rabbits are the framework's interface and its protection and management of checkpoint and data files within the storage hierarchy. On systems without Rabbits, both the interface and the storage hierarchy run within the context of the application on the compute nodes, which means SCR's work could perturb, or slow down, the application's progress. "The real advantage of the Rabbits is that user tools such as SCR can run on the Rabbit nodes, which means these features can run independently of the application, allowing the applications on the compute nodes to immediately return to normal operations and generate results even sooner," says McFadden.

Following the lessons learned in applying SCR to Oak Ridge's and Argonne's exascale computers, the Livermore team is testing SCR and offering it to anyone who wants to streamline I/O checkpointing operations for their applications (SCR is open source and available for download via GitHub). "We want SCR to enable El Capitan to be ready on the first day. We have been running SCR on a series of early software and hardware versions to work out everything we can in advance," says Moody.

Learning and integrating new library systems such as SCR can be a daunting task, but the SCR team hopes more computer scientists in DOE and the National Nuclear Security Administration (NNSA) will apply SCR to new systems. SCR's adaptability makes it a user-friendly and versatile tool for anyone looking to run code on advanced computing systems. Researchers will use El Capitan's immense compute power, remarkable speed, and enhanced simulation capabilities to improve and impact national security missions, and SCR will play a key role in supporting Livermore's mission. Mohror says, "DOE and NNSA both are tasked with doing science that benefits the nation's interests. We've put a lot of work into SCR. The technology has been very successful at Lawrence Livermore, and we think it can help others reduce the time required to gain scientific insight."

—Amy Weldon

For further information contact Kathryn Mohror (925) 423-2997 (mohror1@llnl.gov).