HYPERION RESEARCH

HPC User Forum Update

Arm A64fx and Post-K: Game-Changing CPU & Supercomputer for HPC, Santa Fe, New Mexico

Bob Sorensen
July 2019

IN THIS UPDATE

The HPC User Forum was established in 1999 to promote the health of the global HPC industry and address issues of common concern to users. In April 2019, the 72nd HPC User Forum took place in Santa Fe, New Mexico. This update summarizes a presentation from that meeting entitled*, Arm A64fx and Post-K: Game-Changing CPU & Supercomputer for HPC and its Convergence with Big Data/AI*, presented by Satoshi Matsuoka, Director, Riken Center for Computational Science.

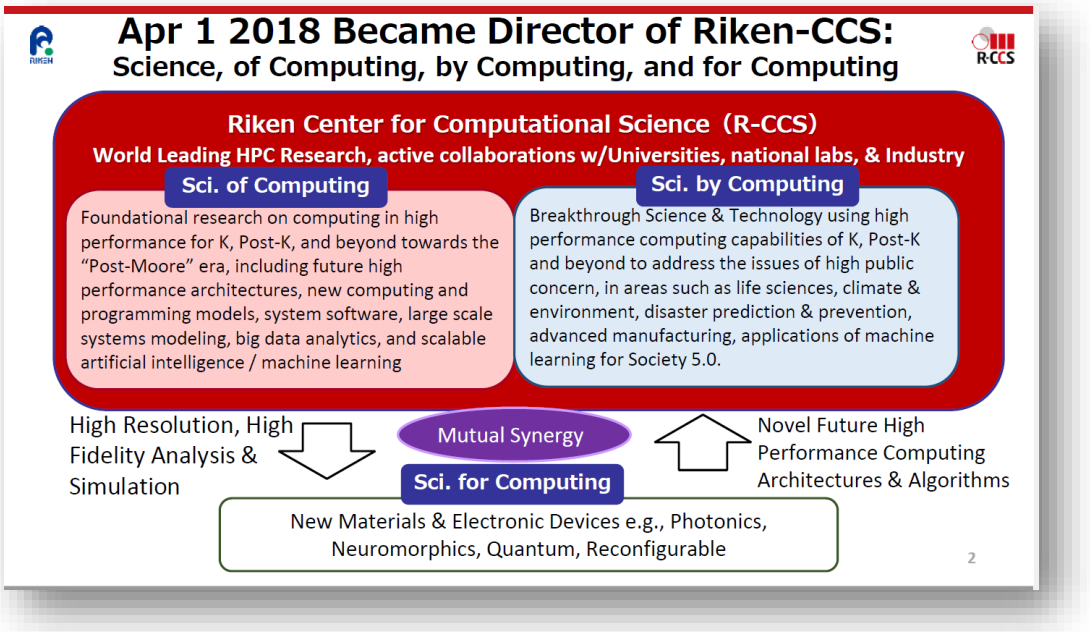
Source: Riken and Hyperion Research, 2019

PRESENTATION: ARM A64FX AND POST-K: GAME-CHANGING CPU & SUPERCOMPUTER FOR HPC AND ITS CONVERGENCE WITH BIG DATA/AI, PRESENTED BY SATOSHI MATSUOKA, DIRECTOR, RIKEN CENTER FOR COMPUTATIONAL SCIENCE

Matsuoka began his presentation with an overview of Riken-CSS, a world-leading HPC research center that actively collaborates with universities, national labs, and the commercial sector. Riken housed and ran the K Computer which was the fastest supercomputer in the world at one time but that is scheduled to be turned off in August 2019. Japanese government officials recently renamed the center RCCS (Riken Center for Computational Science) to highlight that they don't just run the K-Computer or develop post-K.

Riken-CSS conducts a broad range of research on high-end computational science in computer science, the science of computing, and the more traditional computational science, science by computing. Also, Riken-CSS is actively looking at the interactions with other parts of Riken and others to work on breakthrough science including climate and environment, advanced manufacturing, and machine learning. Riken-CSS has also started a project on post-Moore computing in the 2027-2028 timeframe and is also looking at quantum technologies.

FIGURE 1



Source: Hyperion Research, 2019

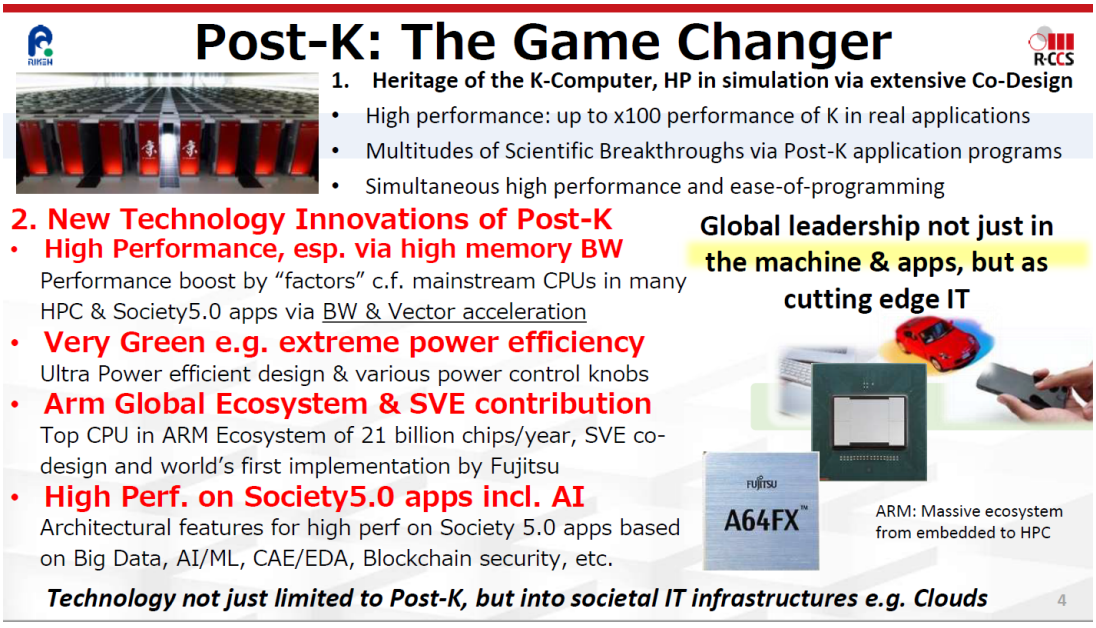
Next, Matsuoka focused on the Post-K and the Arm 64fx chip. The team started the first consideration of the Post-K back in 2009. Officially, the Post-K project started in 2012 as a precursor project. Then, in 2014 it was officially inaugurated. Ultimately, Fujitsu-Riken have been designing this machine for the last 5-10 years depending on how it's measured.

The Post-K system, which was recently officially named Fugaku after Mt. Fuji, the tallest mountain in Japan, will be the largest machine ever built, with more than 150,000 nodes. Matsuoka noted that with the Post-K, it's not as much about flops, but bandwidth. The machine theoretically has more than 150 petabytes per second bandwidth which is only slightly less than an order of magnitude bigger than any other machine in the world for real applications.

- The 6D Torus Network, which is an improved version of Tofu Network on K, gives about 60 petabits of injection bandwidth total. To put that in context, it's about 10x bigger than all the Internet Data Center traffic.
- There is about 25-30PB NVMe L1 storage. And the IO fabric, not the main fabric, is InfiniBand and has 10,000 endpoints.

The Post-K won't hit exascale flops, but Mitsuoka indicated that that's not the goal. The goal is to achieve exascale performance in real applications. In fact, Matsuoka projected that many of the follow-on machines that will claim exaflop will be, by-and-large, equivalent or less-so, at least in parity, with this machine in real applications. But this machine will come out in 2020, at least couple of years earlier than those machines.

FIGURE 2

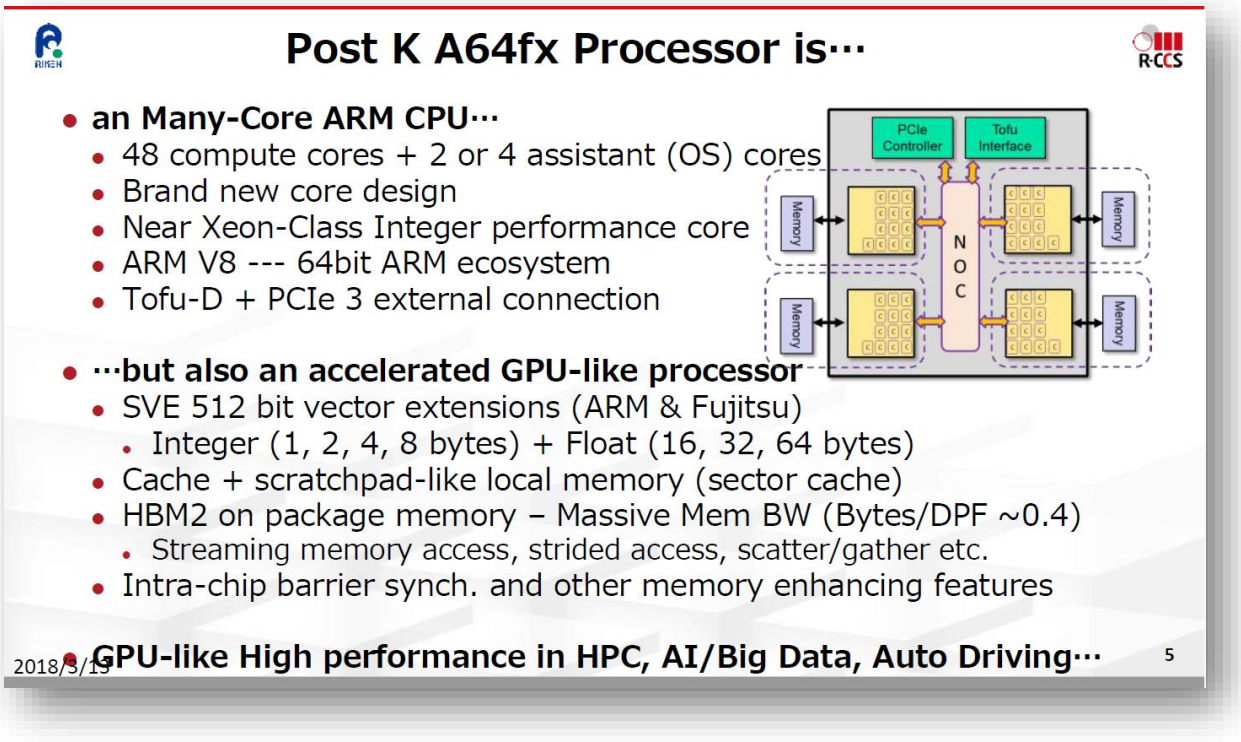


Source: Hyperion Research, 2019

The Post-K uses a Fujitsu custom designed and built processor called the A64fx, 48 core processor build on the 7nm node with a brand-new core design. However, it does not use any of the ARM IP except for the architectural license.

According to Mitsuoka, the ARM instruction set can be used but anything internally within the chip was co-designed by Riken-CSS and Fujitsu. The chip not only has the four quadrants (it's a single chip), but it's a single die with four quadrants. It's all cache coherent. And, in this coherent domain, there is the Tofu Network and the PCIe embedded in the chip. Of note, the chip has HMB2 on package memory supporting a byte per double precision flops ratio of about 0.4.

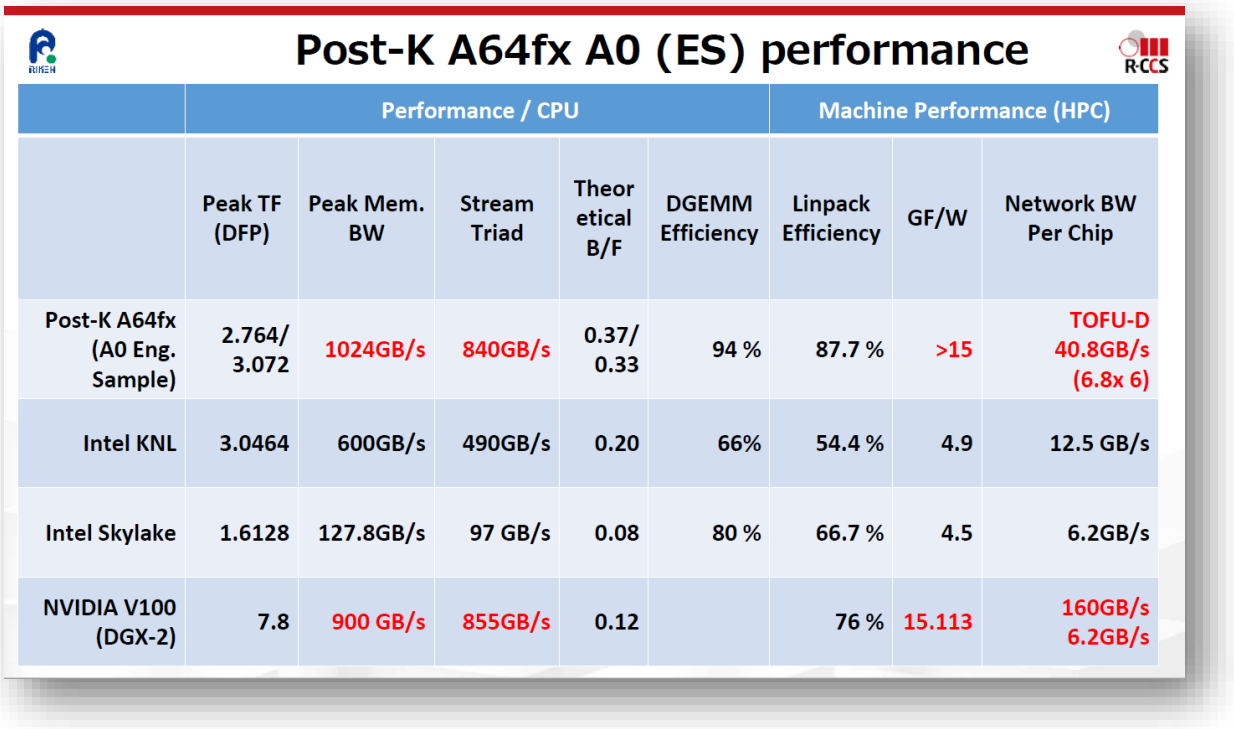
FIGURE 3



Source: Riken and Hyperion Research, 2019

Matsuoka concluded with some performance comparisons of the Post-K with existing HPC counterparts. Based on the data presented, the Post-K offers a world-class peak memory bandwidth, a high bandwidth per flops ratio, impressive GF/W performance that rivals even the best GPU-based system and promises high Linpack efficiency. Plans call for the Post-K, or Fugaku, to move into the assembly phase this year, and for the system to be turned over to early users in early 2020.

FIGURE 4



Post-K A64fx A0 (ES) performance

	Performance / CPU					Machine Performance (HPC)		
	Peak TF (DFP)	Peak Mem. BW	Stream Triad	Theoretical B/F	DGEMM Efficiency	Linpack Efficiency	GF/W	Network BW Per Chip
Post-K A64fx (A0 Eng. Sample)	2.764/ 3.072	1024GB/s	840GB/s	0.37/ 0.33	94 %	87.7 %	>15	TOFU-D 40.8GB/s (6.8x 6)
Intel KNL	3.0464	600GB/s	490GB/s	0.20	66%	54.4 %	4.9	12.5 GB/s
Intel Skylake	1.6128	127.8GB/s	97 GB/s	0.08	80 %	66.7 %	4.5	6.2GB/s
NVIDIA V100 (DGX-2)	7.8	900 GB/s	855GB/s	0.12		76 %	15.113	160GB/s 6.2GB/s

Source: Riken and Hyperion Research, 2019

For more information or to view this and other presentations given at HPC User Forums dating back to 2008, visit www.hpcuserforum.com.

About Hyperion Research, LLC

Hyperion Research provides data driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com
