



OpenFOAM

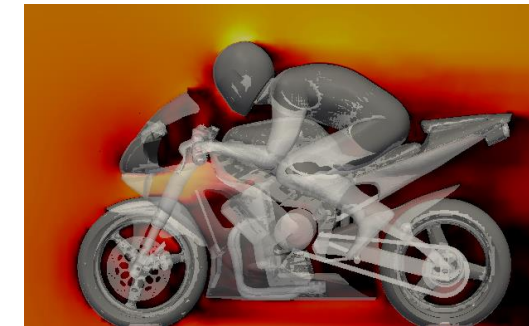
Performance Testing and Profiling

October 2017

- **The following research was performed under the HPC Advisory Council activities**
 - Participating vendors: Huawei, Mellanox
 - Compute resource - HPC Advisory Council Cluster Center
- **The following was done to provide best practices**
 - OpenFOAM performance overview
 - Understanding OpenFOAM communication patterns
 - Ways to increase OpenFOAM productivity
 - MPI libraries comparisons
- **For more info please refer to**
 - <http://www.huawei.com>
 - <http://www.mellanox.com/hpc/>
 - <https://www.openfoam.com/>

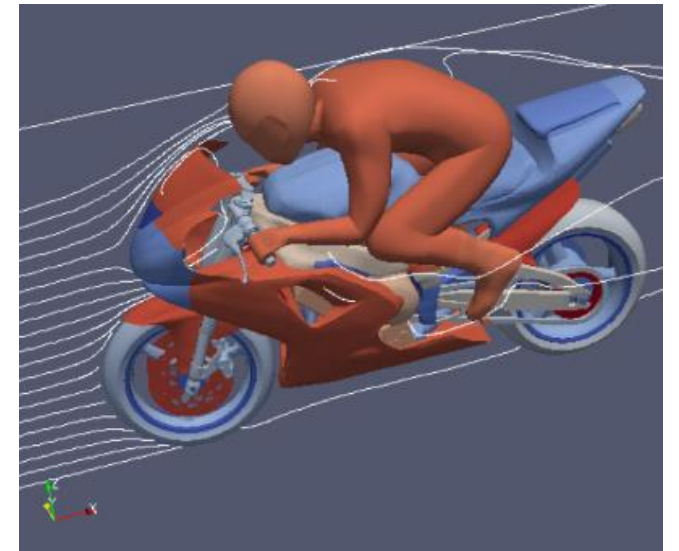
Open  FOAM®

- **OpenFOAM® (Open Field Operation and Manipulation) CFD**
- **Toolbox in an open source CFD applications that can simulate**
 - Complex fluid flows involving
 - Chemical reactions
 - Turbulence
 - Heat transfer
 - Solid dynamics
 - Electromagnetics
 - The pricing of financial options
- **OpenFOAM support can be obtained from OpenCFD Ltd**



- **The presented research was done to provide best practices**
 - OpenFOAM performance benchmarking
 - MPI Library performance comparison
 - Interconnect performance comparison
 - Compilers comparison
 - Optimization tuning
- **The presented results will demonstrate**
 - The scalability of the compute environment/application
 - Considerations for higher productivity and efficiency

- **Huawei FusionServer X6000 “Broadwell” Cluster with Huawei FusionServer XH321 V3 32 Server Nodes**
 - Dual-Socket 14-Core Intel Xeon E5-2690 v4 @ 2.60 GHz CPUs (35MB Cache, Turbo @3.50GHz)
 - Dual-Socket 16-core Intel Xeon E5-2697A v4 @ 2.60 GHz CPUs (40MB Cache, Turbo @3.60GHz)
 - Memory: 256GB memory, DDR4 2400 MHz, Memory Snoop Mode in BIOS sets to Home Snoop
 - OS: RHEL 7.2, MLNX_OFED_LINUX-4.1-1.0.2.0 InfiniBand SW stack
- **Mellanox ConnectX-4 EDR 100Gb/s InfiniBand Adapters**
- **Mellanox Switch-IB SB7800 36-port EDR 100Gb/s InfiniBand Switch**
- **Huawei OceanStor 9000 Scale-out NAS storage system**
- **Compilers: Intel Parallel Studio XE 2018**
- **MPI: Intel MPI 2018, Mellanox HPC-X MPI Toolkit v1.9.7**
- **Application: OpenFOAM v1612+, single precision**
- **MPI Profiler: IPM (from Mellanox HPC-X)**
- **Benchmarks: MotorBike, 160K elements, 100 steps**



About Huawei FusionServer X6000 – XH321 V3

XH321 V3

Entire Server



Node

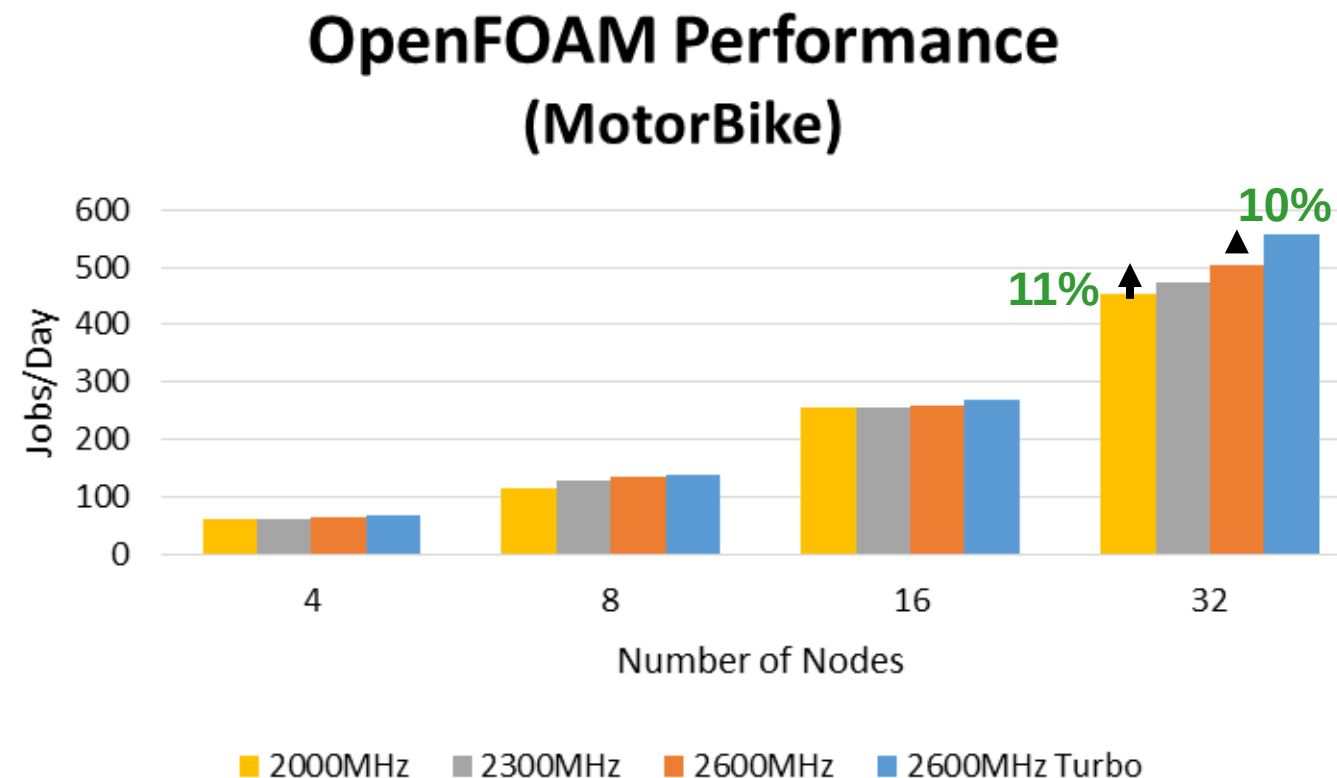


Form Factor	Half-width server node
Processors	1 or 2 Intel® Xeon® E5-2600 v3/v4
Memory	16 DDR4 DIMMs, providing up to 1 TB memory when configured with LRDIMMs
Internal Storage	6 SAS/SATA HDDs or NVMe SSDs
RAID	RAID 0, 1, 10, 5, 50, 6, or 60; supercapacitor
LOM Network Ports	2 GE or 2 GE + 2 x 10GE
PCIe Expansion	Up to 2 PCIe x16 slots Note: The mainboard has two PCIe slots: PCIe slot 1 (left) shared by a RAID controller card and PCIe slot 2(right) shared by an IB card .
LOM Storage	1 SATADOM and 1 M.2 NVMe SSD
Operating Temperature	120 W to 145 W processors: 5°C to 35°C Processors below 120 W: 5°C to 40°C
Dimensions	40.5 mm x 177.9 mm x 545.5 mm (1.59-in. x 7.00-in. x 21.48-in.)

Highlights

- **High performance:** Supports 1 or 2 Intel® Xeon® E5-2600 v3/v4 series processors with up to forty-four cores and 55 MB L3 cache capacity
- **High reliability:** Supports multiple RAID levels, supercapacitor for power failure protection, and TPM disk encryption
- **Diverse LOM I/O ports:** mainboards support up to two 1GbE and two 10GbE LOM ports
- **Features:** Hot-swappable fans, 2.5” NVMe SSDs, Mixed NVMe SSD and SAS/SATA SSD, 1m deep cabinet, 40 °C operating temperature, Liquid Cooling

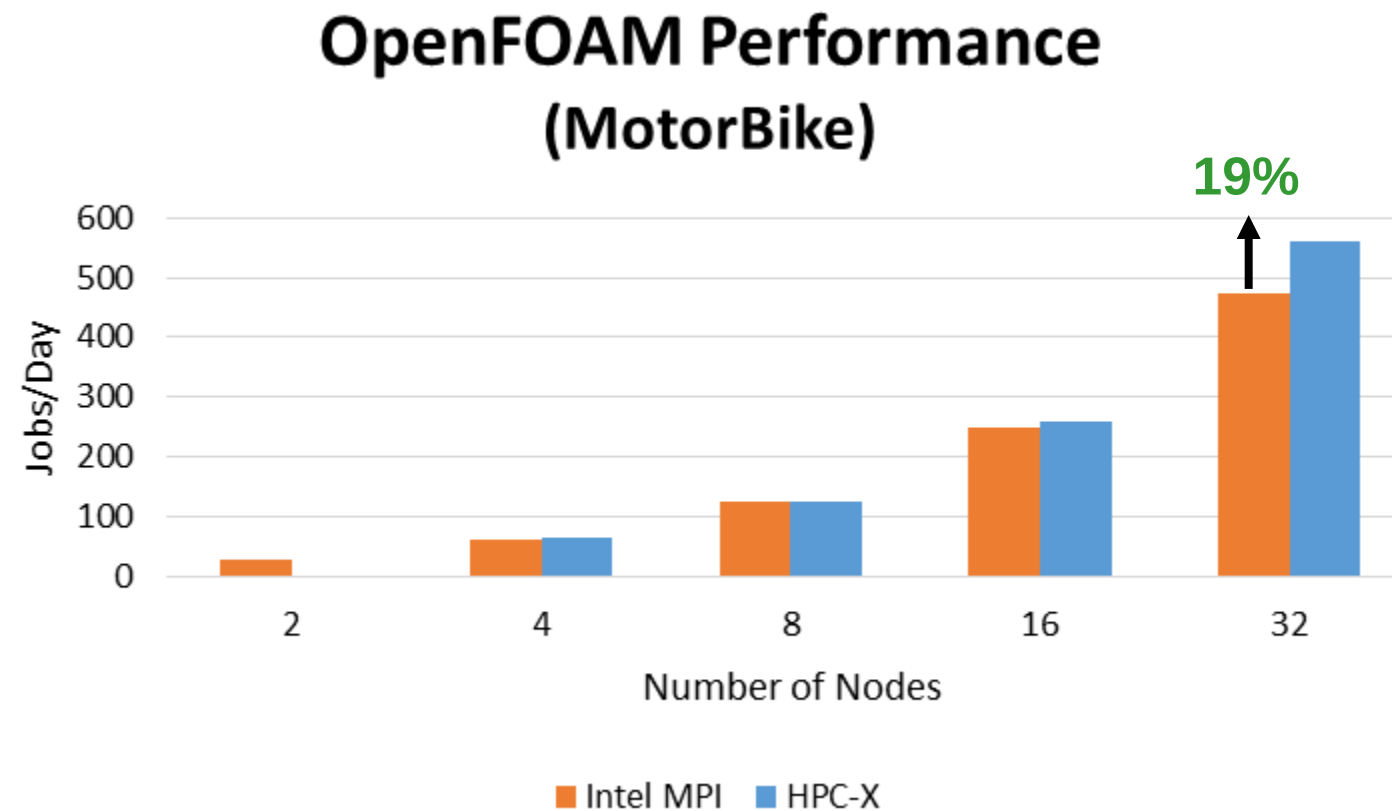
- **OpenFOAM can benefit from higher frequency CPUs**
 - Difference can be ~11% when running 2GHz compared 2.6GHz at base clock (at 32 nodes)
 - Difference can be ~10% when comparing 2.6GHz CPUs with and without turbo speed (at 32 nodes)
 - ‘turbostat’ shows cores runs at turbo speed (3087MHz) when turbo is enabled



Higher is better

28 MPI Processes / Node

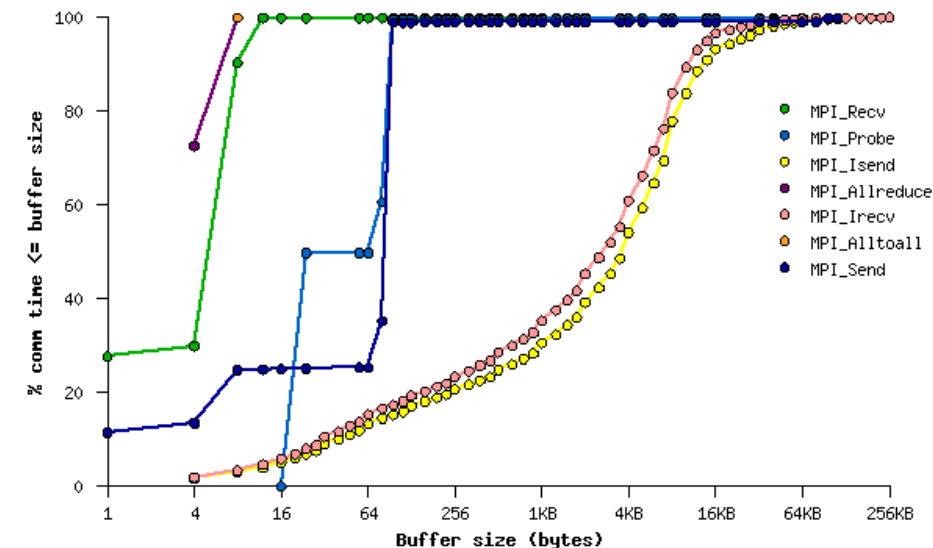
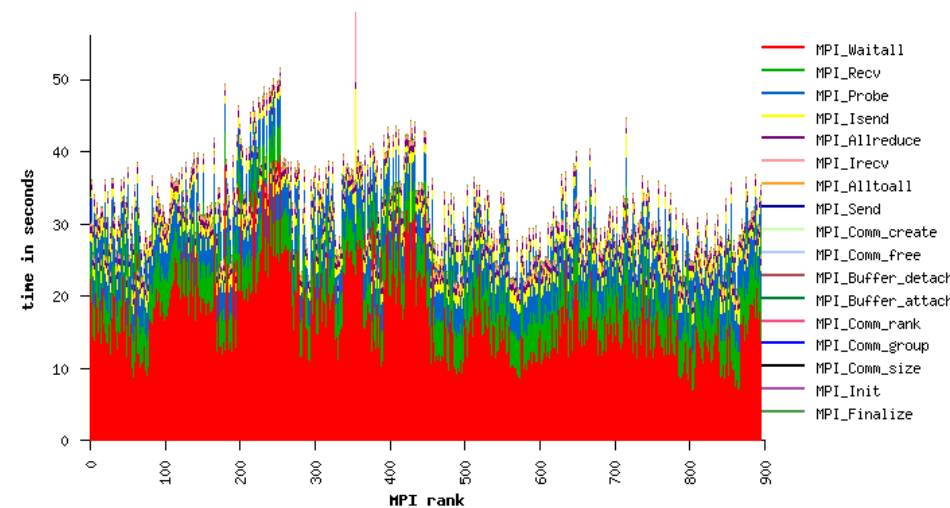
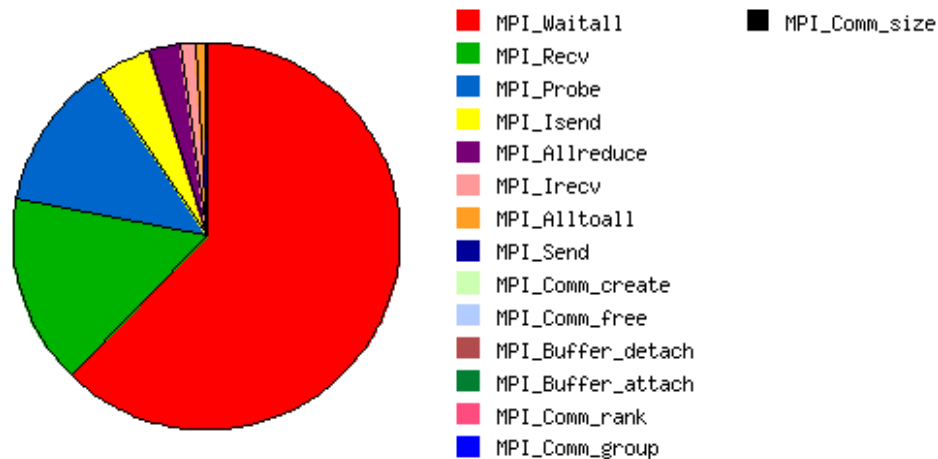
- **Comparing two MPI libraries**
 - HPC-X provides 19% higher performance at 32 nodes



Higher is better

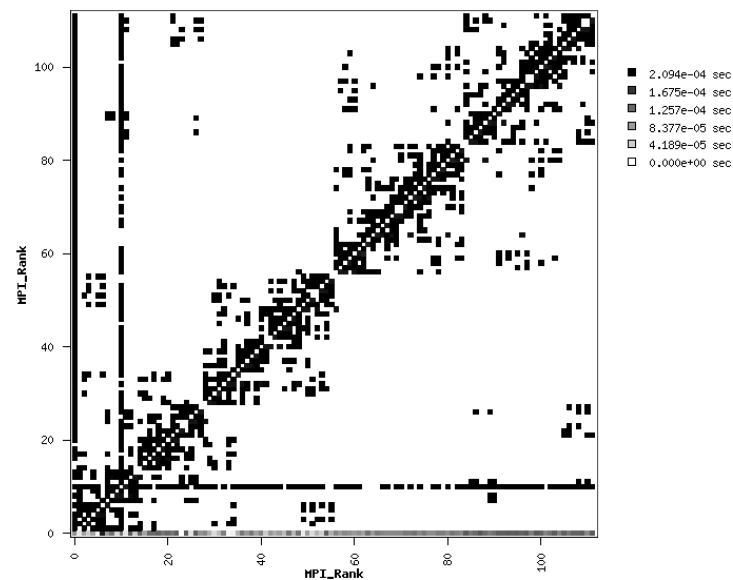
28 MPI Processes / Node

- **MPI profiler shows the type of underlying MPI network communications**
 - Majority of communications occurred are non-blocking communications
- **Majority of the MPI time is spent on non-blocking communications at 32 nodes**
 - MPI_Waitall (11% wall), 8-byte MPI_Recv (1.4% wall), 1-byte MPI_Recv (0.7% wall)
 - Only 14% of the overall runtime is spent on MPI communications at 32-nodes (when EDR is used)

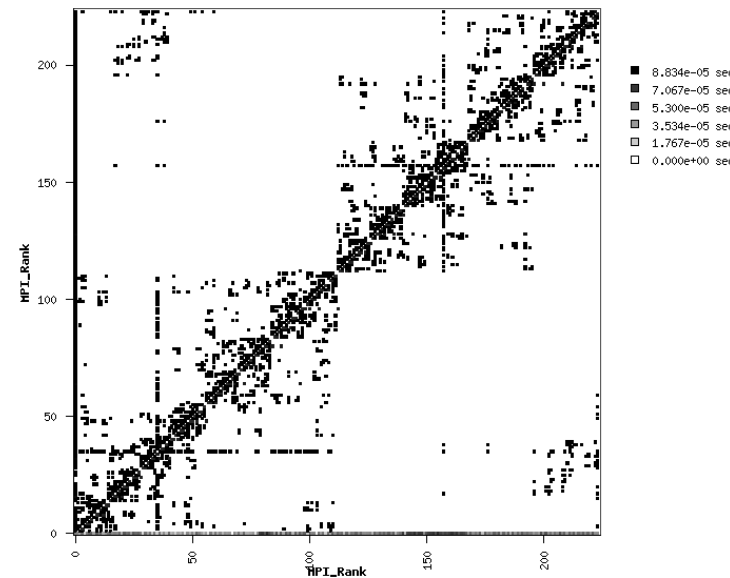


- **Communication topology shows communication patterns among MPI ranks**
 - MPI processes mainly communicates with neighbors, but also shows some other patterns

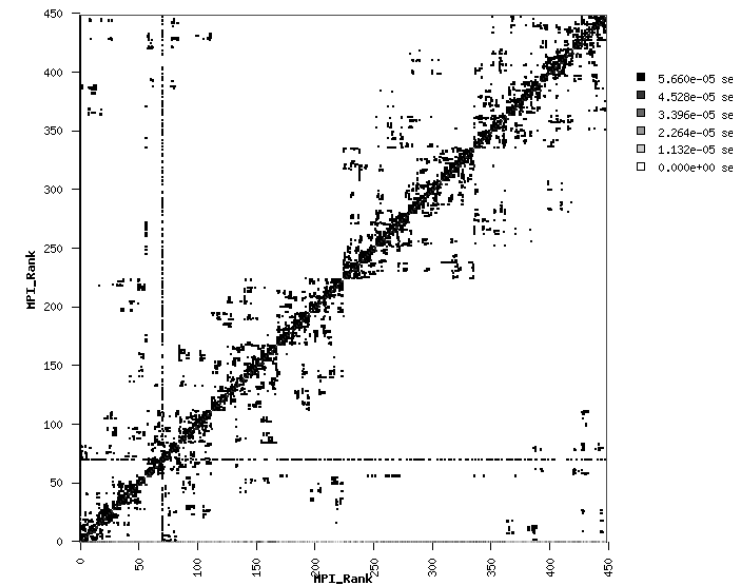
4 Nodes



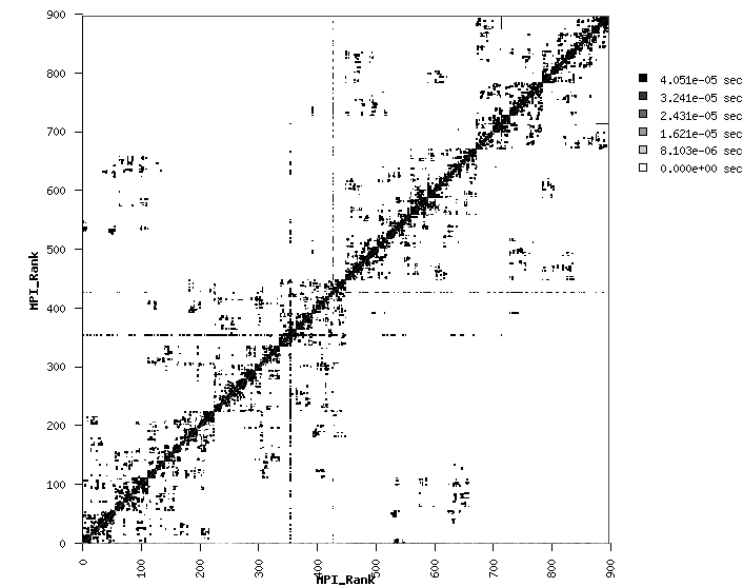
8 Nodes



16 Nodes



32 Nodes



- **Network and MPI library are important factor for better OpenFOAM scalability**
 - HPC-X and tuning can improve performance by up to 19% at 32 nodes
- **OpenFOAM can benefit from CPU that runs at a higher clock speed**
 - ~11% better performance when running 2.6GHz compared 2.0GHz at base clock (at 32 nodes)
 - ~10% better performance when running 2.6GHz CPUs with turbo speed than without (at 32 nodes)
- **MPI Profiling**
 - Most MPI time is spent on MPI non-blocking communications
 - Majority of the MPI time is spent on MPI non-blocking communications at 32 nodes
 - MPI_Waitall (11% wall), 8-byte MPI_Recv (1.4% wall), 1-byte MPI_Recv (0.7% wall)

Thank You

HPC Advisory Council



All trademarks are property of their respective owners. All information is provided “As-Is” without any kind of warranty. The HPC Advisory Council makes no representation to the accuracy and completeness of the information contained herein. HPC Advisory Council undertakes no duty and assumes no obligation to update or correct any information presented herein