CAASCADE: Understanding HPC Applications for Evidence-based Co-design

What's the problem?
Understanding the DOE HPC application landscape is important for effective integration of heterogeneous computational architectures and exploration of future computing technologies. There is a paucity of quantitative data on critical application characteristics, and absence of tools to obtain and manage this data.

Examples:
- Which programming models are being used?
- Which programming model features are critical?
- How important is Fortran to the DOE?

What are we doing now?
The community most often relies on single-use, labor-intensive efforts, "institutional knowledge", or written survey responses and anecdotal input from developers in an attempt to gain insight into diverse and expanding HPC applications.

What should we be doing?
Enabled through CAASCADE

Automate the collection of application program characteristics from a variety of tools.
Create a database to provide access to this information, enabling data analytics and knowledge discovery techniques to inform ongoing HPC research.

Ideal:
- Gather data in a repeatable, normalized (non-human) way
- Compilers and linkers know "everything" about source code
- Convert compiler internal representations to human knowledge
- Understand which application characteristics (e.g. language features, computational motifs) are contributing flops on Leadership systems

Results

Left
High-level information such as Fortran language standard (top left), the type of variable (right). OpenMP Library, a structure parallelism methodologies by number of statements (division left), and subroutine with OpenMP pragmas (division right) in E3SM as collected by CAASCADE.

Below
The distribution of data type characteristics from QMCPACK from the CAASCADE static analysis only (left, and the same data when re-weighted by dynamic runtime information from CrayPAT right). The static only information indicates how the application is being developed and what libraries are contributed flops on Leadership systems.

Below
Libraries and the number of calls sites detected with XALT and CAASCADE in NUCOR. Libraries depicted in red belong to LAPACK and in brown belong to HDF5.

Below
Summary information about OpenMP usage in the E3SM application. The usage of specific OpenMP statements is shown (left), and the proportion of code covered within OpenMP lexical extents is shown per subroutine (right). Only the top fifteen are shown for clarity.

Below
MRF routines used in the E3SM application. The frequency of calls for each MRF routine is shown. Only the fifteen most called routines are shown for clarity.