FFT, FMM, and Multigrid on the Road to Exascale
Performance Challenges and Opportunities

Extended Abstract

Huda Ibeid
University of Illinois at Urbana-Champaign
Urbana, Illinois
hibeid@illinois.edu

Luke Olson
University of Illinois at Urbana-Champaign
Urbana, Illinois
lukeo@illinois.edu

William Gropp
University of Illinois at Urbana-Champaign
Urbana, Illinois
wgropp@illinois.edu

ABSTRACT
FFT, FMM, and multigrid methods are widely used fast and highly scalable solvers for elliptic PDEs. However, emerging systems are introducing challenges in comparison to current petascale computers. The International Exascale Software Project Roadmap [6] identifies several constraints on the design of exascale software. Challenges include massive concurrency, energy efficiency, resilience management, exploiting the high performance of heterogeneous systems, and utilizing the deeper and more complex memory hierarchy expected at exascale. In this study, we perform model-based comparison of the FFT, FMM, and multigrid methods in the context of these constraints and use the performance models to offer predictions about the methods performance on possible exascale system configurations, based on current technology trends.

1 CONCURRENCY
To gain some insight into the solvers performance on the massively concurrent systems expected at exascale, we derive analytical performance models that include computation and both intra- and inter-node communication costs. The intra-node communication along with the computation cost account for the single node performance which is a critical building-block in scalable parallel programs, whereas the inter-node term reflects the impact of network communication on the scalability.

Assuming arithmetic and memory operations can be overlapped, the total execution time $T_{\text{exe}}$ is given by

$$T_{\text{exe}} \approx \max(T_{\text{comp}}, T_{\text{mem}}),$$  \hspace{1cm} (1)

where $T_{\text{comp}}$ is the computation time and $T_{\text{mem}}$ is the time spent transferring data in a two-level memory hierarchy between the main memory and a cache with size $Z$ and cache-line length $L$ in elements. The inter-node communication cost is modeled using the $(\alpha, \beta)$ model where a message send cost can be represented as

$$T_{\text{net}} = m \alpha + n \beta,$$  \hspace{1cm} (2)

where $\alpha$ represents communication latency, $\beta$ is the send time per element over the network, $m$ is the number of messages sent, and $n$ is the total message size.

The Roofline model can be used to assess the quality of attained floating-point performance (GFLOP/s) by combining machine peak performance, machine bandwidth, and arithmetic intensity. Figure 1 shows a roofline model of an AMD Istanbul processor along with the arithmetic intensity of various phases of FFT, FMM, and multigrid.

2 ENERGY
Power is a major hurdle on the road to exascale. Achieving an exaFLOP will be limited to no more than 20 MW of power when an exascale system could use as much as 100 MW if existing petascale design methodologies are used. This imposes design constraints on both the hardware and software to improve the overall efficiency.

To compare power and energy efficiency of the FFT, FMM, and multigrid methods, we use the energy Roofline model introduced in [4]. The energy Roofline model defines energy cost (Joules) as

$$E \equiv E_{\text{flop}} + E_{\text{mem}} + E_0(T_{\text{exe}}),$$  \hspace{1cm} (3)

where $E_{\text{flop}}$ is the total energy of computation, $E_{\text{mem}}$ is the total energy of memory operations, and $E_0$ is the total constant energy.
To estimate the energy efficiency on future multicore chips, we use the transistor scaling projection model presented in [7]. Figure 2 shows how the energy efficiency is predicted to improve over time as transistor sizes decrease. Nevertheless, the per-transistor energy efficiency improvements have slowed dramatically in comparison to the historic rates. Microarchitecture innovations are needed to achieve the power budget imposed for exascale machines.

3 RESILIENCY

The exponential increase in core counts expected at exascale will lead to increases in the number of routers, switches, interconnects, and memory systems. Consequently, resilience is a major challenge for HPC applications on future exascale systems.

To asses resiliency, we quantify the vulnerability of the data structures and communication schemes. For the data structure, we use the data vulnerability factor (DVF) introduced in [11]. The DVF for a specific data structure (DVF$_d$) is defined as

$$DVF_d = FIT \times T_{exe} \times S_d \times N_{ha},$$

where $FIT$ is the failure in time (failures per billion hours per Mbit), $S_d$ is the size of the data structure, and $N_{ha}$ is the number of accesses to the hardware (memory) in this study.

Figure 3 shows the DVF of the FFT, FMM, and multigrid methods. The results show that algorithms with random memory access pattern, such as the FMM, have higher DVF (more sensitive to memory errors) in comparison to algorithms with template-based memory access patterns, such as FFT and multigrid.

For communication, we introduce the communication vulnerability factor (CVF) which reflects the communication scheme as well as the network characteristics. The CVF for a specific kernel (CVF$_k$) is defined as

$$CVF_k = m \times T_{net} \times R_n,$$

where $R_n$ is the network resilience.

Figure 4 shows that the all-to-all communication scheme of FFT makes it more sensitive to network failures in comparison to the hierarchical methods, FMM and multigrid. The hierarchical nature of FMM and multigrid reduces the communication complexity of FFT to $O(\log P)$ communication complexity.

4 HETEROGENEITY

As accelerators advance in both performance and energy efficiency, heterogeneous systems have become critical ingredient in the pursuit of exascale computing. One bottleneck to consider on heterogeneous systems is the PCIe bus. Due to the high compute capability of the GPU, the PCIe bus can have a significant impact on the performance. The PCIe transfer time for $n$ elements is given by

$$T_{PCIE}(n) = n \beta_{PCIE},$$

where $\beta_{PCIE}$ is the I/O bus inverse bandwidth in seconds per element.

Figure 5 shows memory and network access costs of the FFT, FMM, and multigrid methods. We consider extrapolated GPU- and CPU-based systems that have the same peak performance. The results show that all methods spend less time in both forms of communication on the CPU-based system. However, this system requires over 10× as many processors as the GPU-based system. This cost could be prohibitive.

ACKNOWLEDGMENTS

This material is based in part upon work supported by the Department of Energy, National Nuclear Security Administration, under Award Number DE-NA0002374.
REFERENCES


