# Roofline Performance Model for Distributed Memory Architectures

Boakye Dankwa

May 5, 2016

## 1 Abstract

The roofline performance model[2] is a simple insightful tool that bounds kernel performance on multi-core systems. In this report, an attempt is made to extend the roofline performance models to a 3-D roofline model for a homogeneous distributed memory systems, taking into consideration the peak floating point performance per node, the peak memory throughput per node and the peak communication throughput. A 3-D roofline performance model is constructed for 16 AMD Opteron Interlagos x86_64 nodes on BigRed II and used to bound the performance of the ***SUMMA*** algorithm [1] on BigRed II.

## 2 Introduction

The roofline performance model bounds kernel performance on multi-core shared-memory systems. It is a 2-D visual performance tool that can provide useful insight into kernel performance and optimization on a given system. It defines a performance upper bound on a given system as:

$$\begin{aligned} Attainable\,GFlops/sec &= min\{PeakFloatingPointPerformance, \\ &\qquad PeakMemoryBandwidth \times OperationalIntensity\}. \end{aligned} \tag{1}$$

which relates the kernel performance, system parameters and kernel operational intensity. While the model can provide useful insight on shared-memory systems, it does not account for kernel communications between nodes in a distributed memory system hence cannot be applied to such systems. In this project, it is attempted to extend the roofline performance model to a class of distributed memory systems such as clusters. The roofline model is extended to a 3-D visual performance model which relates system parameters, network parameters, kernel memory and communication traffic.

# 3 Previous Work

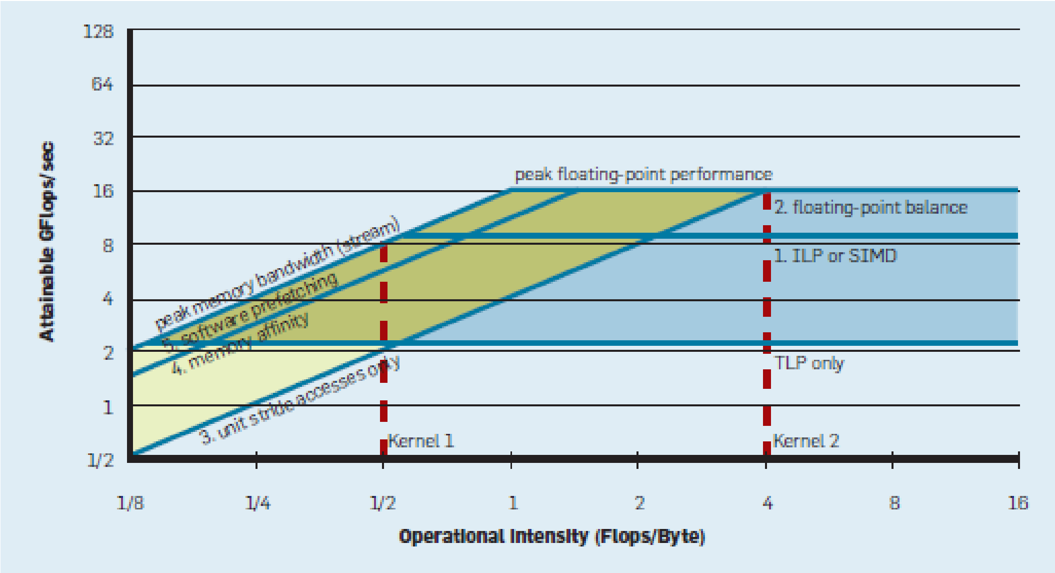

Figure 1: Roofline Model for Multicore Systems

Patterson and Co.[2] related the performance of a given kernel with the peak floating point performance of the system, the peak memory bandwidth beyond the system cache, and kernel operational intensity. They first define the obtainable performance on a system as in equaton 1. That is, the peak performance that can be obtained on a given multiprocessor system is upper bounded by the minimum of the peak floating point performance, a property of the system, and the kernel's performance constrained by the system's memory bandwidth. They call this upper bound the roofline. To provide insight into kernel performance, the author's added computational and memory bottleneck bounds to the model. These bounds, called ceilings, divides the 2-D diagram, shown in Figure 1, into regions and provide insight into how a kernel may be modified to improve performance. They explained that a given optimization technique must be performed to break through its bottleneck ceiling on the model. The computational bottleneck ceilings include thread-level parallelism (TPL) optimization ceiling, instruction-level parallelism (ILP) optimization ceiling and data-level parallelism (SIMD) optimization. These performance ceilings are defined in equations 2, 3 and 4 respectively. The constants in these equations obtained from the system's data sheet.

$$\begin{aligned} TLP\,bound &= number\,of\,cores \times frequency && (2)\\ ILP\,bound &= number\,of\,cores \times frequency \\ &\quad \times max\left(1, \frac{ThreadPerCore}{Latency}\right) && (3)\\ SIMD\,bound &= number\,of\,cores \times frequency \\ &\quad \times (SIMD\,width/SIMD\,throughput) && (4) \end{aligned}$$

The memory bottleneck ceilings include the unity memory stride ceiling, which bounds the performance if the kernel accesses memory contiguously, memory affinity ceiling, which bounds the performance if each processor employs memory affinity and also accesses only local memory, and software prefetching ceiling which bounds performance if kernel combines unity stride access, memory affinity and software prefetching. The authors obtained the memory ceilings using a benchmarks such as STREAM.

The author's demonstrated the model on four multi-core systems: Intel Xeon (Clovertown, e5345), AMD Opetron X4 (Barcelona 2356), Sun UltraSPARC T2+ (Niagrara 2, 5120) and IBM Cell (QS20), and several floating point kernels including SpMV, LBMHD, Stencil and 3D FFT. They concluded that their models provided good insights into the performance of the kernels on these multicore systems, especially the ridge-points- the minimum operational intensity to achieve maximum performance - proved to be better performance predictors than the clock rate of peak performance.

# 4 The 3-D Roofline Model

The 2-D roofline performance model applies to shared memory multi-core systems. The model cannot be effectively applied to distributed memory systems. In such systems, considerations must be given to peak network bandwidth as well as kernel *communication intensity* on the system. We define *communication intensity* as the number of floating point operations per word of network traffic. Therefore, *communication intensity* gives an indication of network bandwidth consumption of a given kernel on a given distributed memory system.

## 4.1 Communication Model

Assuming that all nodes resides on a LAN, let's choose a simple communication model on the systems.

| MPU Type | AMD Opteron 16-Core | NVIDIA Kepler GK110 GPU |
|---|---|---|
| **ISA** | x86/64 | x86/64 |
| **Total Threads** | TBD | TBD |
| **Total Cores** | TBD | TBD |
| **Total Sockets** | TBD | TBD |
| **GHz** | TBD | TBD |
| **Peak GFlops/sec** | TBD | TBD |
| **Stream GB/Sec** | TBD | TBD |

Table 1: Characteristics of two multi-core systems

| Kernel | Operational Intensity | Description |
|---|---|---|
| **dgemm** | TBD | BLAS kernel for $C = C + AB$ |
| **SpMV** | TBD | TBD |

Table 2: Characteristics of two multi-core systems

## 5 Results

## 6 Conclusion and Future Work

## 7 Acknowledgements

## References

[1] Jerell Watts R. A. Van De Geijn. Summa: Scalable universal matrix multiplication algorithm. , 1996.

[2] D. Patterson S. Williams, A. Waterman. Roofline: An insightful performance model for multicore architectures. *COMMUNICATIONS OF THE ACM*, 2009.

| System | Nodes Used | Cores/Node | Peak Flops/s | Peak Stream Memory Bandwidth | Peak Communication Bandwidth |
|---|---|---|---|---|---|
| **Big Red II** | 16 | 1 | 160GFlops/s | 22.8GBytes/s | 4MBytes/s |

Table 3: Characteristics of distributed memory systems

| Kernel | Operational Intensity | Communication Intensity | Description |
|---|---|---|---|
| **SUMMA** | 32 | 2048 | Parallel matrix multiplication |

Table 4: Kernel to demonstrate 3-D Roofline model