Cross-architectural Modelling of Power Consumption Using Neural Networks

Miloš Puzović[1], Eun Kyung Lee[2], Vadim Elisseev[3]

[1]The Hartree Centre, [2]IBM T.J. Watson Research Center, [3]IBM Research

Main Contributions

- We extend recent work [1, 2] on estimation of power consumption of HPC systems with metrics obtained using hardware performance counters to three different micro-architecture implementations: Intel 64 Broadwell, IBM POWER8 and Cavium ThunderX ARMv8 architecture,
- We argue that this methodology is portable across different micro-architecture implementations.
- We discuss the optimal number and type of hardware performance counters required to accurately predict power consumption within few percents of the actual power consumption.
- We improve accuracy of power consumption predictions by employing a Neural Networks (NN) based model

Motivation

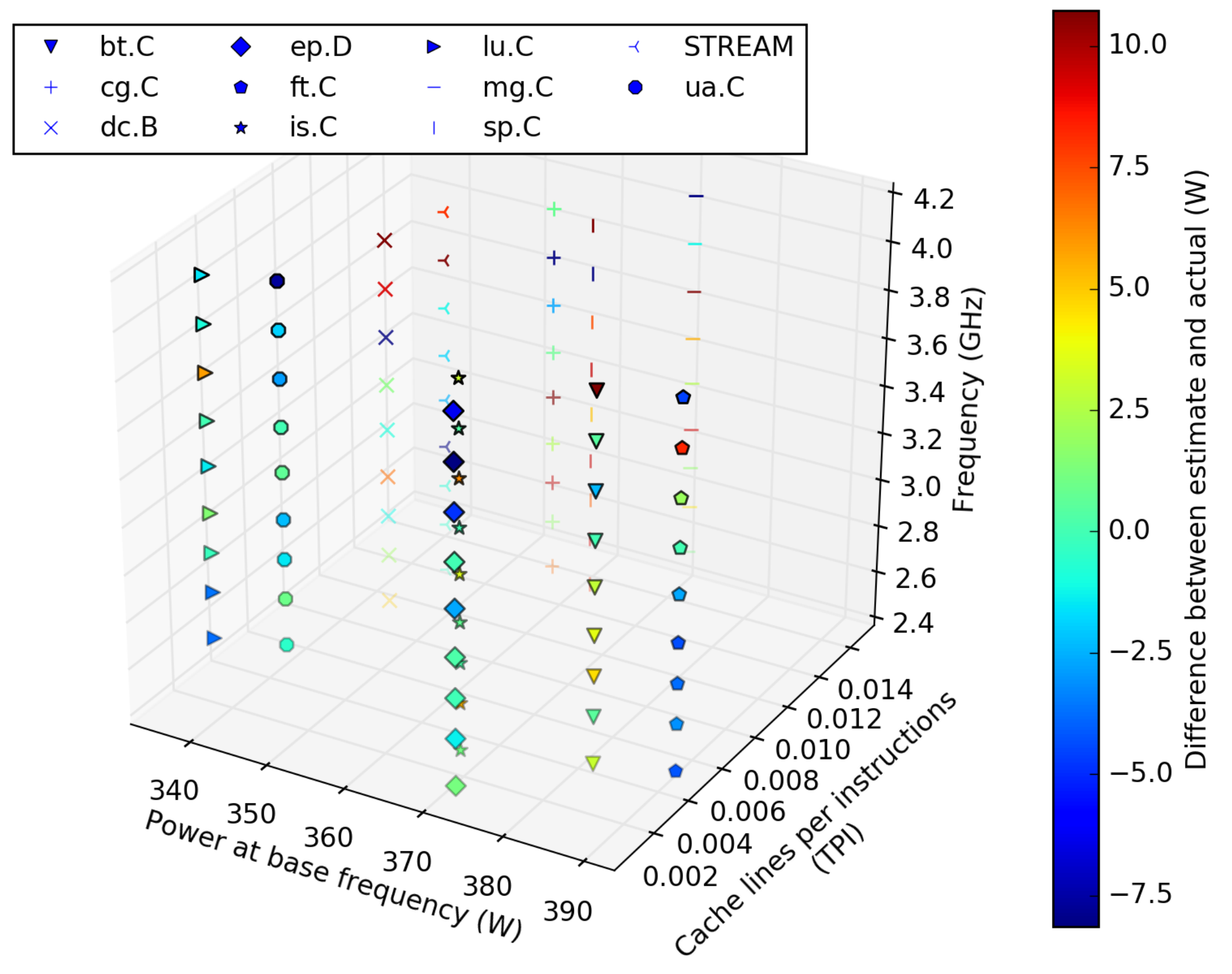


Figure 1: Accuracy of power consumption estimation

Prior models as per Figure 1:

① overestimate or underestimate large number of benchmarks with different frequency to nominal

② do not cover the whole spectrum of possible combinations

③ power consumption averaged over application run-time

Architectures

Architecture		IBM POWER	Intel x86-64	ARMv8 64bit
Processor		Power S822LC	Intel Xeon E5-2698 v4	Cavium ThunderX
Core	Frequency	3.5 GHz	2.3GHz	2.0GHz
	# of cores	10	16	48
	# of threads	80	32	48
Execution unit	Type	out-of-order	out-of-order	in-order
	# of issue/commit	10 / 8	8 / 4	2 / 2
L1D Cache	Policy	NUCA	Write-allocate	Write-through
	Type	Private	Private	Private
	Size	64 KB/core	32 KB	32KB
	Associativity	8-way	8-way	32-way
L1I Cache	Size	32KB/core	32KB	78 KB
	Associativity	8-way	8-way	39-way
L2 Cache	Policy	NUCA	Write-back	Write-back
	Type	Private	Private	Shared
	Size	512KB/core	256KB	16MB
	Associativity	8-way	8-way	16-way
L3 Cache	Policy	NUCA	Write-back	N/A
	Size	8MB/core	40MB	N/A
	Type	Shared	Shared	N/A
SMP Interconnect	Bus Type	SMP	QPI	CCPI
	Bus speed	9.6GB/s per channel	9.6GB/s	10.3GHz
Memory	Type	DDR4 1600	DDR4 2133	DDR4 2133
	# of channels	8	4	4
	Access speed	1600 MHz	2133 MHz	2100 MHz

Table 1: Three different microarchitecture implementations

Neural Network (NN)

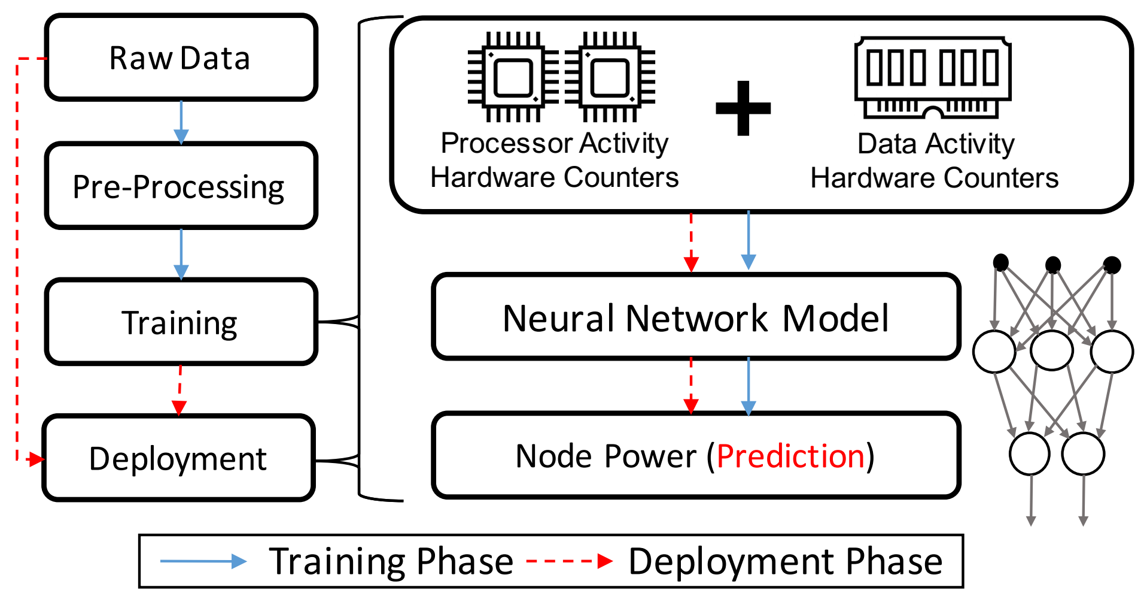


Figure 2: Neural Network-based prediction approach.

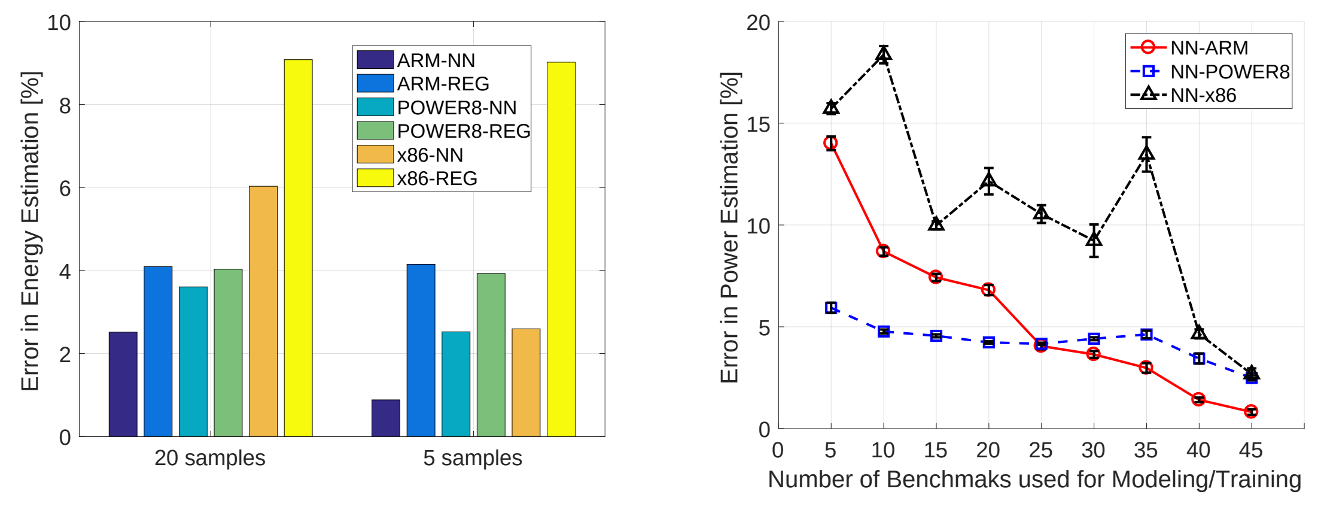
Event	Intel Xeon E5 v4	IBM S822LC	Cavium ThunderX
IPC	EVENT_CPU_CLK_UNHALTED EVENT_INST_RETIRED	PM_RUN_CYC PM_RUN_INST_CMPL	CPU_CYCLES INST_RETIRED
IFETCH	EVENT_ISSUED	PM_INST_DISP	ISSUE
STALL	EVENT_RESOURCE_STALLS	PMU_CMPLU_STALL	STALL_BACKEND
BR	EVENT_BR_INST_EXEC EVENT_BR_MISP_EXEC	PM_BR_CMPL PM_BM_MPRED_CMPL	BR_RETIRED BR_MIS_RETIRED
FLOPS	FP_ARITH_INST	PM_FLOP	ASE_SPEC VFP_SPEC
L1	MEM_LOAD_RETIRED_L1_HIT	PM_DATA_FROM_L2	L1D_CACHE_REFILL L1D_CACHE
LCCM	UNC_CBO_CCACHE_LOOKUP.ANY_REQ UNC_CBO_CCACHE_LOOKUP.I UNC_ARB_TRK_REQUEST.EVICTIONS	PM_MEM_READ PM_MEM_PREF PM_MEM_RWITM	L2D_CACHE_REFILL_LD L2D_CACHE_REFULL_ST L2D_CACHE_WB_VICTIM L2D_CACHE_WB_CLEAN

Table 2: Hardware performance counters used for raw data

The input layer consists of processor activity hardware counters and data activity hardware counters as shown in Table 2, which we have identified to be the major sources of power draw. An entire data set using five benchmark suites was collected for three different micro-architectures as shown in Table 1.

Accuracy

We have experimented with different number of runs per each benchmark to test impact of data set size on the model accuracy as illustrated in Figure 3:



(a) Each benchmark run (b) Total benchmarks

Figure 3: NN-model power estimation accuracy

We have noted that if the same benchmarks is run multiple times, the overall accuracy of NN model increases and that accuracy of the NN model increases as we increase number of different benchmarks in the training set. We have noticed no further increases in accuracy beyond 20 for each benchmark run which indicates a convergence threshold for the NN model.

Predictors

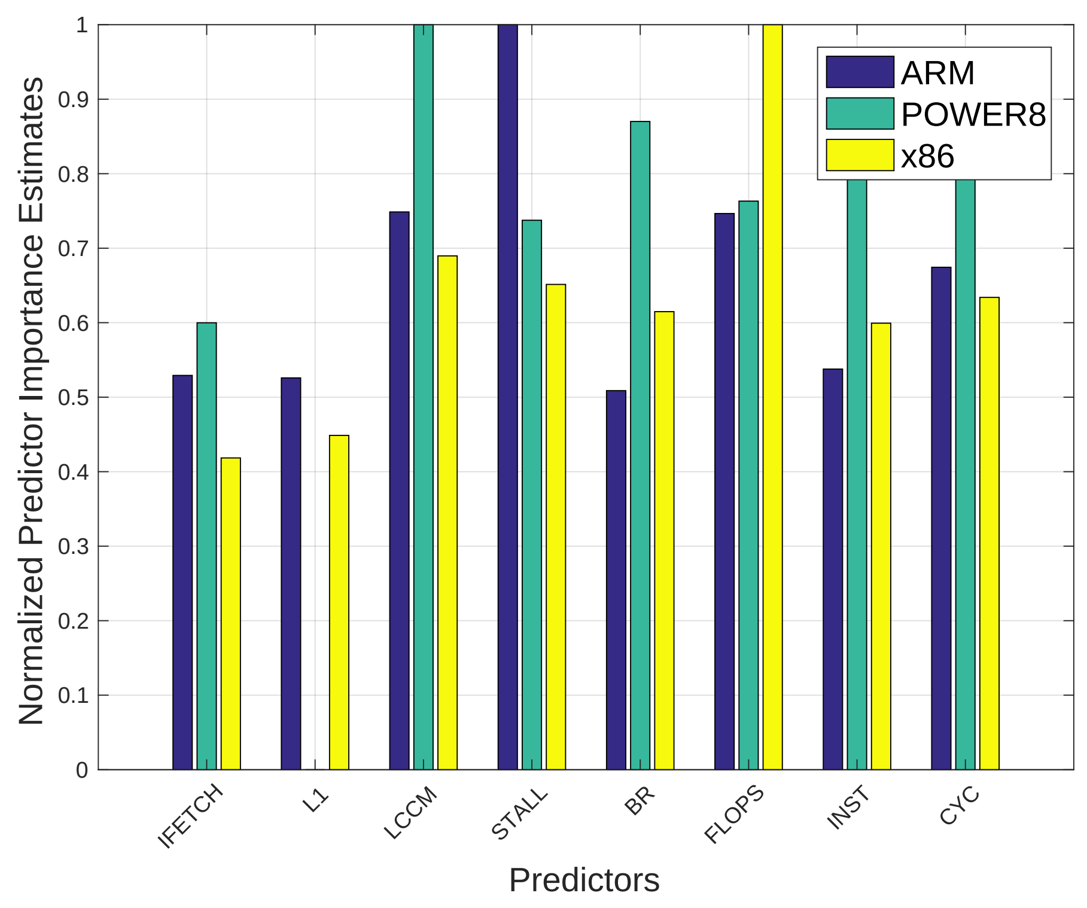
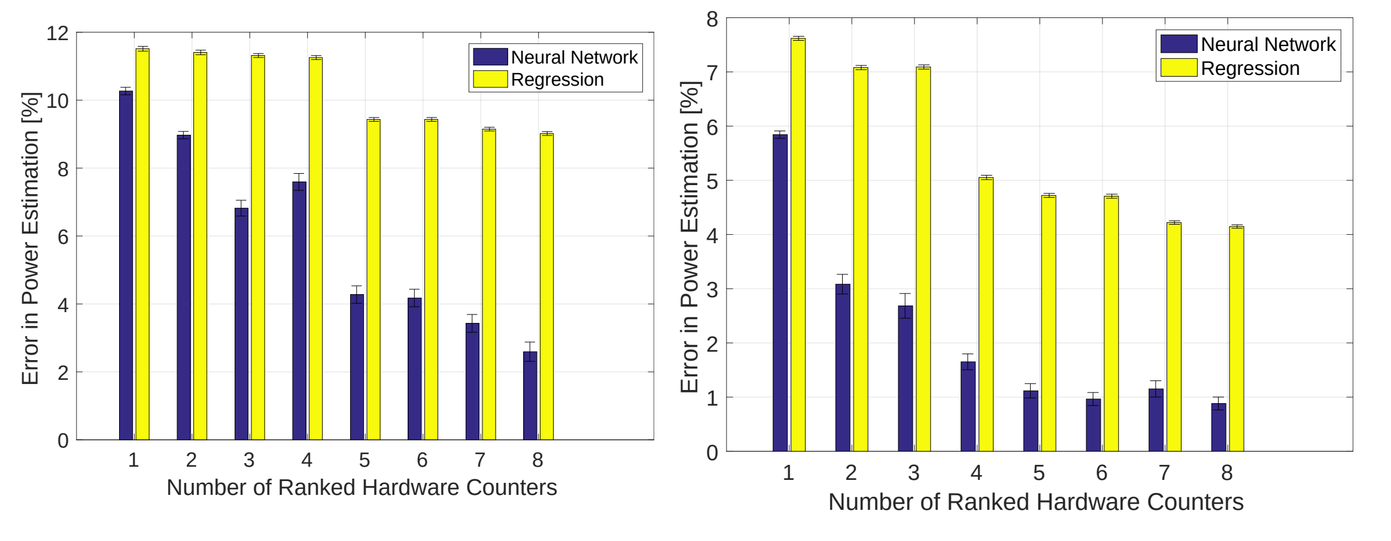


Figure 4: Importance of each hardware performance counters in estimating power consumption using neural network model

Figure 4 suggests that it would be very difficult to use neural network model developed for one architecture to estimate power consumption on the other architecture. Importance of different hardware counters for estimating power consumption is different for different micro-architectures.

Results

The Figures below show for each microarchitecture error in power estimation between regression and neural network model.



(a) Intel 64 (b) Cavium ThunderX

Figure 5: NN-model power estimation error

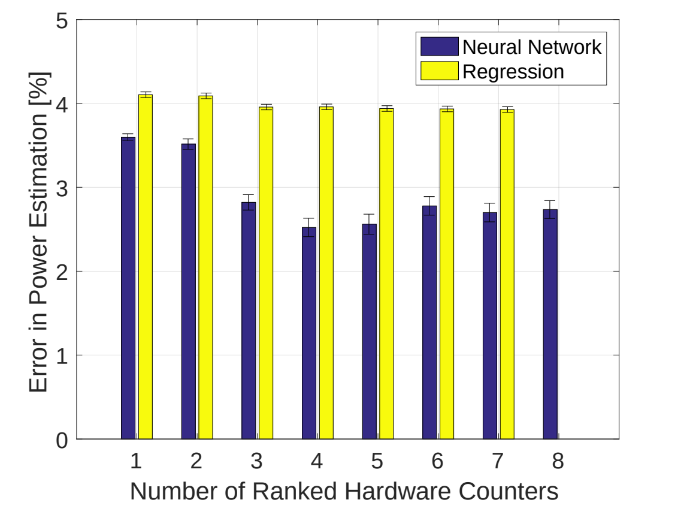


Figure 6: IBM Power8 NN-model power estimation error

In order to have the same basis for the comparison between neural network and regression model we have updated the regression model from to take into account hardware performance counters that we have sampled as per Table 2.

Microarchitecture	IBM POWER8 S822LC	Cavium ThunderX ARM
Error (%)	77%	64%

Above Table shows results when model trained on Intel 64 is applied to to estimate power on Cavium ThunderX and IBM Power 8.As expected since Intel64 and Cavium ThunderX give similar importance to the same hardware performance counters the power estimation on ThunderX is more accurate then on IBM Power8. Unfortunately, the accuracy is significantly worse when compared to Figures 5 and 6 and this is mainly due to difference in importance of the same hardware performance counters that each microarchitecture implementation assigns.

References

[1] Auweter et al.
A Case Study of Energy Aware Scheduling on SuperMUC.
In *ISC 2014*, pages 394–409.

[2] Elisseev et al.
Energy Aware Scheduling Study on BlueWonder.
In *4th E2SC@SC16*, pages 61–68.