End-to-end Modeling and Optimization of Power Consumption in HPC Interconnects

(Invited Paper)

Sébastien Rumley, Robert P. Polster, Keren Bergman
Lightwave Research Laboratory
Department of Electrical Engineering
Columbia University, New York, NY, USA
rumley@ee.columbia.edu

Simon D. Hammond, Arun F. Rodrigues
Scalable Computer Architecture
Sandia National Laboratories
Albuquerque, NM, USA

***Abstract*— The Interconnect topology is one of the key design choices of large-scale distributed computer architectures. It might also become one of the most power consuming design elements as traffic volumes and interconnect size continue to grow. High interconnect power consumption can be simply provoked by non-energy efficient components, or can in contrast be due to architectural misconception. In this paper, we propose and combine various high-level models to realize a clear breakdown of the power consumptions, and analyze how these depend on various parameters, either external or internal, to the interconnect. Our initial results indicate that end-to-end interconnect consumption is dominated by routers. The node compute power can also affect the interconnect energy efficiency, especially if links of equal bandwidth are used as injection links and topology inner links.**

Keywords—Interconnection networks, network topology, energy efficiency.

I. INTRODUCTION

Large scale computing systems rely on massive parallelization. To reconcile partial computations realized in a myriad of independent computer nodes, a high-speed interconnection system is required. Interconnects providing connectivity to tens of thousands of end-points, and allowing each end-point to inject and receive multiple tens of Gigabit/s, have been developed for the most recent supercomputers. At such scales, ensuring a rapid delivery of communication messages at all times and between all source-destination pairs is a big challenge. This is especially true under tight cost, but also, and increasingly, power constraints.

The selection of the topology determines by a large extent the cost, power consumption, and performance of an interconnect. It has thus received a sustained attention from the research community [1-3]. Highly related to the topology selection is the number of ports available on each router, called router radix r, and the total number of nodes (determining the number of end-points N in the interconnect) of the envisioned supercomputer. Going through n r-ported routers allows reaching at most r^n distinct destinations, it is thus easy to see that $r^n > N$, and thus that $n > \log_r N$ must hold. Hence, the overall size of the topology is fundamentally determined by the logarithm in base r of the number of end-points N. Going through n routers also means *occupying* a router n times, and *transiting* over $n+1$ links. Consequently, the number of routers and links required, independent of their number, scales with $\log_r N$ at least.

For a fixed N, $\log_r N$ decreases as r grows. Larger radixes are thus a synonym of smaller numbers of links and routers, making routers with the largest port counts the most interesting. The router's inner complexity, however, scales with r^2 [4], and too large port counts are infeasible in a cost effective die area. The power dissipation of routers, as further detailed later, can be expected to scale superlinearly with the number of ports as well, setting another upper bound to scalability. Finally, pinout constraints must also be taken into account. Therefore, r is limited to values clearly smaller than N. In practice, r is of the order of one hundred while N is of the order of 10,000. A fundamental aspect of topology design consists therefore in finding an advantageous trade-off between router complexity, a function of r, and global interconnect size, a function of n.

Next to the number of end-points and the router radix, the computing capabilities of each compute node and the bitrate of interconnect links play an important role in the topology selection process, too. Naturally, higher performing nodes can inject higher amounts of traffic into the interconnect. The bitrate offered by the links connecting them to the interconnect (node-to-router links - NR) must thus be dimensioned in proportion. As for interconnect inner links (router-to-router - RR), their bitrate can be similar to the one of NR links, in which case router ports can be arbitrarily used for either NR or RR links, an advantage. However, designs with distinct bitrates on NR and RR links have also been proposed. This allows selecting the bitrate according to the respective distance and technology of NR and RR links. As the longest links in large scale supercomputer often exceed 10 meters, optical transmission is generally required for RR links. Provided the high bitrate capabilities of optics, it may look advantageous to tolerate higher bitrates on RR links than on NR ones. In this way, traffic is funneled in less numerous RR links. However, high bitrate RR links impact the router complexity, too, which may introduce limits on the radix side. Reducing r increases n,

due to the logarithmic scaling rule. This might negate the benefits of high bitrates in the RR links.

The design of a large scale topology is therefore a subtle alchemy in which the router radix, the number of end-point and RR and NR bitrates are all interrelated. Other practical concerns that should be considered include: embedding servers and routers in cabinets and rows; ease of mapping of an application's traffic pattern on the topology [5-6]; capacity to support strongly typed traffic patterns [7].

In such complex design spaces, early decisions have to be taken. In particular, the selection of the switch radix and of the NR and RR bitrates is often dictated by commercial availability and cost, following a market analysis [3][8]. In that case, the variables to be adjusted are either the topology shape, if the number of end-point N is part of the specifications, or N. N must be adjusted if a specific topology, advantageous in terms of cost, bi-section bandwidth or number of hops, but imposing a number of clients, is targeted [9]. Having the NR bitrate dictated also imposes to some extent limitations on the node compute power (expressed in FLoating point OPeration per Second - FLOPS).

Fixing one or more interconnect characteristics as working assumptions and determining the remaining ones as a function of the earlier is generally sub-optimal. As the largest supercomputers try to break the Exascale barrier, such sub-optimal decisions must be addressed. This holds in particular for the issue of power consumption. The US Department of Energy, among the world's largest users of ultra-high scale machines, has fixed a design target of 20 MW for future supercomputer platforms. To achieve an ExaFLOPS within this budget, architectures delivering 50 GigaFLOP per Joule (or, equivalently, 50 GFLOP/s/W) must be designed. This is 7 times better than the most power efficient HPC system today [10] and generally 25 times more efficient than the majority of recent systems, including the world's most powerful HPC system [11]. Tremendous efforts are thus required to achieve 50 GFLOP/J while scaling beyond 1 ExaFLOPS. On the interconnect side, reaching the required energy efficiency implies a per bit end-to-end power budget of ~20 pJ/bit. This means squeezing all interconnect related operations into the power budget allocated, today, to a single link [12].

This observation has motivated us to consider the topology design problem from a more abstract but also more holistic angle, and to elaborate a methodology allowing critical design variables to "float". Our objective is to find in which proportions the bitrate of NR and RR links, r, N and the node computer power in FLOPS should be selected to minimize the end-to-end power consumption of the interconnect. To reach this objective, we first leverage the approach previously reported in [1], which describes the relationships between the number of links, the number of end-points N and the radix r in ideally connected and balanced topologies (Section II). We also define two power models permitting to relate the consumption of board-level electrical links and optical links to the bitrate. A third power model is introduced to relate the router consumption to its bi-section bandwidth (Section III). Our models are then combined and applied to various supercomputer designs. The system-level design space exploration is conducted along three axes: relative strength of compute nodes in FLOPS (which dictates the number of nodes necessary to reach a global computing power); concentration of compute nodes around switches (which dictates the number of routers) and imbalance between bitrates of NR and RR links.

We show that routers are by far the highest contributors to end-to-end power consumption. To minimize router consumption, maximal router bandwidth envelopes need to be utilized. Our router model estimates the router bandwidth envelopes of state of the art routers to be 6Tb/s. As power consumption of links grow with their bitrate, it is preferable to divide router bandwidth over more links. This influences the choice of processing node compute power, unless imbalance between NR and NN links is introduced. We discuss these results further in Section V and conclude.

II. Lower Bounds for Balanced Topologies

Our approach presented in [1] permits to bound the requirements (in terms of router radix, number of routers, and number of RR links) for interconnecting N end-points. This approach applies to *direct* topologies exclusively. In direct topologies, all routers are equivalent and receive the same number of NR links. In indirect topologies, routers are associated with a level or a stage, and receiving NR links is exclusive at stage 0. As our topology model is limited to direct topologies, only those are considered in this paper.

We define C as the number of NR links each router is attached to (C is called the concentration factor), and R as the number of RR links attached to each router. We immediately can observe that all interconnect requirements can be expressed in terms of C and R (and N). Hence, the number of routers in the interconnect is given by $S = ceil(N/C)$, the number of links is given by $S \cdot R$ and the required radix r is $C+R$. The challenge resides in deciding how to optimally split r among C and R. Obviously C must be equal or greater to 1, while R must be equal or greater than 2. If R=1, routers can only be connected in pairs which would result in a disconnected topology.

With the minimal value of R=2, routers can be connected following a ring topology. This minimal, sparse connectivity result in average distance between routers of $\Delta = S/4$. In contrast, with $R = S-1$, routers are connected in an all-to-all fashion. In that maximally connected case, the average distance separating two routers is Δ=1. The average distance between routers in an interconnect Δ is an important metric. It dictates how many times in average a packet (or single bit) is emitted by routers and transmitted over RR links, thus how many times it consumes resources. Suppose every end-point sends one unit of traffic to the router it is attached to, and that this traffic is evenly destined to all other end-point (all-to-all uniform traffic). Each of these units of traffic will be repeated Δ times before reaching the egress router. The interconnect must thus be capable of transporting $N\Delta$ units of traffic to support this traffic pattern.

Suppose now that each RR link can support one unit of traffic. Since $S \cdot R$ links are present, $S \cdot R \geq N \cdot \Delta$ must hold. As $N = S \cdot C$,

$$\Delta \leq R/C = R/(r-R). \quad (1)$$

In other terms, in order for a topology to be able to absorb a uniform traffic pattern at the bitrate of the RR links, the R/C ratio must be superior to the average distance.

Δ depends on S, R, but also on how the $S \cdot R$ links are distributed across the router pairs. Without the knowledge of the topology structure, it may appear difficult to decide on the C - R splitting. However, we can assume that topologies employed in HPC systems generally favor low average distances, in particular if they need to support applications inducing traffic over a large portion of the node pairs. The lower bound for the average distance of a graph with S vertices and degree R is the average distance of an hypothetical Generalized Moore Graph (GMG) and is given by

$$\Delta_{GMG}(R,S) = \frac{R + 2R(R-1) + 3R(R-1)^2 + \ldots + Dx}{S-1} \quad (2)$$

where D is the resulting diameter and x is a remainder. As shown in our prior work [1], topologies showing a Δ relatively close to the bound can be identified. Therefore, $\Delta_{GMG}(R,S)$ can be used to verify if the R/C ratio is compliant with Eq. (1).

If a radix r is imposed, the smallest value of R for which Eq. (1) holds can be found by successively testing candidates in increasing order. Here, we remark that with Δ being at least one, we obtain $R/(r\text{-}R) \geq 1 \equiv R \geq r/2$. Candidate enumeration can therefore be initiated at $R = r/2$. By contrast, if r is not imposed, a pair of (C,R) values can be found for each concentration factor C, theoretically until $C = N$. In that particular case, all nodes are attached to one switch (S=1), no RR links are required, and $C = r$. In practice, C can be limited to a third of the maximal envisioned radix, e.g. ~50. Table I shows various balanced (C,R) pairs along with corresponding Δ_{GMG}, $S \cdot R$ and r values. We notice that for low radixes, C and R are rather imbalanced as many RR links are required to carry the higher traffic induced by the large number of hops. For N=10,000, a minimal radix of six is required. This minimal radix grows with N. For $N=32^3$=32,768 the minimal radix is 7.

TABLE I. EXAMPLES OF BALANCED DESIGNS

N=10,000				N=32,768			
(C,R)	Δ_{GMG}	$S \cdot R$	r	**(C,R)**	Δ_{GMG}	$S \cdot R$	r
(1,6)	5.41	60,000	7	(1,7)	5.60	~230k	8
(2,9)	3.85	45,000	11	(2,10)	4.43	~163k	12
(5,15)	2.88	30,000	20	(5,17)	3.24	~111k	22
(10,24)	2.40	24,000	34	(10,28)	2.75	91,756	38
(15,30)	1.95	20,010	45	(15,36)	2.39	78,660	51
(20,39)	1.92	19,500	59	(20,40)	2.00	65,560	60
(25,47)	1.88	18,800	72	(30,59)	1.94	64,487	89
(30,55)	1.83	18,370	85	(40,77)	1.9	63,140	117
(40,69)	1.72	17,250	109	(50,93)	1.86	61,008	143

A. Injection bandwidth and NR/RR link bitrates

We consider the bandwidth required on a node-router link (NR link) to be equal to the product of the node computing power ($\Pi_{compute}$ - in FLOPS) with the workload verbosity to be supported (v - in byte/FLOP). Note that unit of the product of the two terms is byte/s. The workload verbosity factor v relates the amount of data, sent over the interconnect, by a workload (in bytes) to the number of compute operations realized (in FLOP) by this workload. If for instance each compute node delivers 3 TeraFLOPS (TF) and a workload verbosity of 0.01 byte/FLOP must be supported, the NR link bandwidth is 30 GB/s. We remark that dimensioning the NR bandwidth below this value would restrict the supported verbosity level, while provisioning above would remain widely without effect.

The bitrate of RR-links, by contrast, is not subject to direct constraints and can be selected equal, larger or smaller to the NR rate. We therefore introduce a parameter κ representing the ratio between bitrates or NR (ρ_{NR}) and RR (ρ_{RR}) links. Normalized to the bitrate of RR links, the total traffic injected becomes $N \cdot \Delta \cdot \kappa$, while the interconnect "capacity" remains $S \cdot R$. Eq. (1) thus becomes

$$\Delta \leq R/\kappa C = R/\kappa(r\text{-}R) \quad (3)$$

III. POWER MODELS

Based on Eq. (2) and (3), we can obtain the properties of various balanced topologies for N and κ; i.e. the required radix ($r = C+R$) and the number of RR links ($S \cdot R$). If we further assume the compute power of each node $\Pi_{compute}$ and the targeted workload verbosity factor v to be known, we can calculate the rate of NR links ρ_{NR}, thus the rate of RR links $\rho_{RR} = \kappa\rho_{NR}$. Thus, we know all characteristics of links and routers utilized, as well as their quantities.

In this section, we aim at translating these characteristics into power consumption figures, to calculate later a gross interconnect wide power consumption. We introduce two power models relating the consumption of short (electrical) and long (optical) distance links to the bitrate, as well as a model relating the consumption of a router to its number of ports and line-rate.

We immediately underline that all these elements are subject of constant improvement and (potentially break through) innovation. It is therefore nearly impossible to create transistor-level based power consumption models which would cover the entire design space. Another approach consists in, first, detailing the main physical processes involved in the different sub-parts (e.g. transmission [13], optical modulation [14], switching [15] or buffering), second, in analyzing their respective relationship with scaling factors, and third, in determining lower bound consumptions for each atomic operation. This analytical modeling approach is extremely useful to delimit the scaling potential of each component. However, these scaling models (big O notation) neglect any constants. These constants can play a major role if an optimum is expected at the lower part of the scales which limits the applicability of this approach.

Alternatively, power consumption figures can be extracted from the knowledge of the community, by collecting data points reported in the literature or in commercial product data-sheets. While this approach accurately captures the constants ignored by the analytical approach, it also captures the trade-off decisions taken by their designers, which may result in a flawed design for a particular figure-of-merit (here, power). Furthermore, scientific articles often report the behavior of a subcomponent only, while data-sheets do not disclose internal parameters. This makes cross-comparison across results particularly difficult.

No approach being totally satisfactory for our purpose, we combined elements of the two paths (analytical and review-based) to develop our models. These should be taken as educated, best-effort guesses. We assume them to be accurate enough to support the demonstration of our end-to-end interconnect optimization approach, but not sufficiently detailed and validated to take any quantitative decision based on these results.

A. Highspeed electrical transceivers

For modeling the consumption of chip-to-chip lanes, we mainly rely on literature results. The gathered database is shown Fig. 1. Each data point represents the tuple of per channel bitrate and power consumption expressed as energy per bit. To be coherent throughout our model, we included only publications that used either a 65 nm, 45 nm or 40 nm CMOS technology. Furthermore, we normalized the data points to bitrate per (differentially signaling) channel, to incorporate pin limitations stemming from the router chips.

Finally, we use an affine approximation to model the database, as we are not aware of first order reasons to consider a more complex model. The fit is also shown in Fig. 1. For,

$$E_{ELEC}(B_{CHANNEL}) = \alpha_{ELEC} \cdot B_{CHANNEL} + \beta_{ELEC} \quad (4)$$

where E_{ELEC} is in energy per bit and $B_{CHANNEL}$ the bitrate per channel, we obtain α_{ELEC} = 189 fJ/bit/Gbps and β_{ELEC} = 1496 fJ/bit.

It is important to note that this model returns the efficiency of a *channel*. To obtain the energy efficiency of a *link* as function of B_{LINK}, we need first to obtain the bitrate per channel. We get the bitrate per channel by dividing B_{LINK} by the number of channels available, itself depending on the total pin allocated to this link. For example, if B_{LINK} = 40Gb/s and 16 pins are available, 4 channels at $B_{CHANNEL}$= 10Gb/s can be used (2 differential channels are required for bi-directional links, thus 4 pins). The resulting link will show an efficiency of E_{ELEC}(10Gb/s) = 3.38 pJ/bit. For any rate B_{LINK}, the efficiency is bounded by β_{ELEC} and approaches this value when the number of pins is very large (in which case, $B_{CHANNEL}$ tends to zero).

By assuming that the efficiency of a link is equal to the efficiency of its channels, we also assume no correlation between efficiency and number of lanes. As in practice parallel lanes can share components (e.g. the clock data recovery), a more sophisticated model taking this correlation into account could also be envisioned. This would permit to amortize a part of β_{ELEC} across channels and yield to better efficiencies. Thus, our model is rather conservative.

B. Optical link

Several publications of VCSEL based optical links report power efficiencies between 1 to 1.5 pJ/bit [16-19]. These values are achieved for bitrates ranging from 10 to 25 Gbps, without showing a clear trend. Our recent analyzes of silicon photonic devices also allow us to forecast pJ/bit links in the near future [20-21]. In this context, we simply assume a bitrate independent energy efficiency for the optical links of 1 pJ/bit.

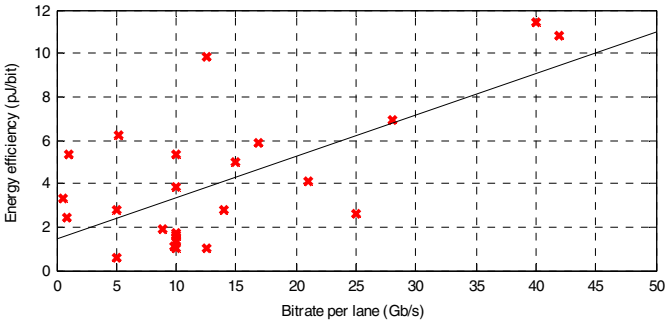

Fig. 1. Efficiency of 22 recently proposed transceiver design.

Note that this efficiency, however, corresponds to the optical segment only. Commercial transceivers and active optical cable embed additionally circuitries to receive and emit high-speed electrical signals on the electrical side. An energy-per-bit expense as calculated in previous sub-section (of at least 1.496 pJ/bit) must thus be added to the "optical" expense. This suggests that the consumption of an optical transceiver is dominated by the electrical part.

C. Modeling router chips

The microarchitecture of a packet router has been discussed in several papers [4], [22-24]. As illustrated in Reference 22, a router chip is composed of, first, electrical transceiver blocks (IO blocks – IOB), in charge of adapting the external bitrate and link format to a standard digital logic signaling form, and second, of a switch core [24], responsible for dispatching packets to the right output port. Router chips are generally air cooled [22] thus limited by their Thermal Design Power (TDP) in the order of 100W [25]. This maximal thermal envelope is shared among the IOBs and the switching core. Routers are also limited by pinout constraints. For instance, Binkert *et al.* [22] consider 1280 IO pins as the upper-limit, an assumption that we make as well. Provided that each port requires 4 pins, this limits the radix to 320.

Our modeling of a router chip, comprising r ports each supporting a bitrate B (in the two directions), begins by allocating the pin budget across the ports. Since we focus on optimizing power (as opposed to area or cost), we divide the maximum pin count possible PIN_{max}=1280 by the number of ports to obtain the number of pins per port $PIN_{max}/r = PIN_{port}$. We further divide PIN_{port} by $4 \cdot B$ to obtain a normalized rate per bi-directional channel $B_{CHANNEL}$, which we introduce in Eq. (4). The resulting energy efficiency is multiplied by $B \cdot r$ to obtain the total IOBs power.

We now estimate the energy efficiency of the switch core. We gathered specifications from commercial products, and selected the most power efficient designs of InfiniBand QDR, FDR and EDR routers. Their specifications are listed in Table II. To extract the consumption of their switching cores, we first divide the listed powers by a power supply efficiency factor 70% as suggested by [26]. 70% efficiency can be considered as rather low for power optimized designs, however, we assume it to include all other overheads (e.g. monitoring LEDs, the management terminal, etc.). We then apply our electrical transceiver model to estimate the IOBs power efficiency. Assuming 10, 14 and 25G per pin, the efficiency of QDR, FDR

and EDR lanes is estimated to be 4.25, 5.11 and 7.48 pJ/bit respectively. These efficiencies, multiplied by $B \cdot r$, are further deducted from the router chip powers to obtain switching core power estimations. These are listed in Table III (last column) along with IOBs, Core, and over-all per-bit energy efficiencies.

We note that power consumption can be assumed "largely independent" of the radix for constant total bandwidth [4]. This holds provided that the power required for arbitration is negligible [27] compared to the power associated with buffering and cross-bar configuration. Hence, we assume the router consumption to scale with its total bandwidth, i.e. with $B \cdot r$. Fitting core powers of Table III leads to a power efficiency expression

$$P_{LINEAR} = \gamma_{CORE} + \delta_{CORE}\, rB \qquad (5)$$

with γ_{CORE} =50.68W and δ_{CORE} = 8.15W/Tb/s (Fig. 2). Deviations of this model compared to core consumptions of Table III remain within 17% difference.

As reported by Passas et *al.* [23], in some designs the power consumption scales quadratically (or more) with r. Superlinear scaling with r can also be observed in switch core estimations provided by Binkert *et al.* [24]. We therefore investigated a second model expressing consumption as

$$P_{QUAD} = \varepsilon_{CORE} + \zeta_{CORE}\, rB + \eta_{CORE}\, r^2 \qquad (6)$$

This "quadratic" model (as opposed to the "linear" model) is showing the least deviation when η_{CORE}=0.08mW/port^2 , ε_{CORE} = 50.64W and ζ_{CORE} = 8.06W/Tb/s. We note that both ε_{CORE} and ζ_{CORE} are very close to γ_{CORE} and δ_{CORE} and that η_{CORE} is very small. The two models (linear and quadratic) are thus very similar. Under the quadratic model, power consumption is substantially impacted by radix only for $r >$ 250, for which the third term of Eq. (6) is 5W. Considering that the radix is limited to 320, we opted for the use of the linear model exclusively.

Figure 3 shows the "feasibility region" under a power dissipation constraint of 132W (maximum value of "Chip power" column in Table III) and pin constraint of 1280. At maximal radix r=320 the largest supported rate is 19 Gb/s, for a total bandwidth of 6.08 Tb/s. On the other side of the Pareto-front, 6, 7 and 8 ports routers support line-rates of 1Tb/s, 870 Gb/s and 765 Gb/s for total bandwidth of 6, 6.1 and 6.12 Tb/s. As shown in Fig. 4a, power constraint designs are all limited to ~6 Tb/s. This is expected, as both core and transceiver consumptions scale with total bandwidth. This also holds for transceivers because each pin carries more bitrate. Figure 3 also shows percentages of the core consumption relative to router consumption, which range from 71.3% to 99.9%. Apparent discontinuities correspond to changes in number of lanes per port (e.g. one lane with radix 161 but two with radix 160). We finally look at the global energy efficiency figures. To obtain the total router consumption, we reapply a 70% efficiency factor to the total power (transceivers and core), to later divide this total power by $B \cdot r$. For power constraints designs with total bandwidth of ~6Tb/s, the power efficiency is ~ (132W / 0.7)/6 Tb/s =~ 31.4W/Tb/s = 31.4 pJ/bit. If total bandwidth is not maximized, power efficiencies raise as γ_{CORE} is amortized over less bits (Fig. 4b).

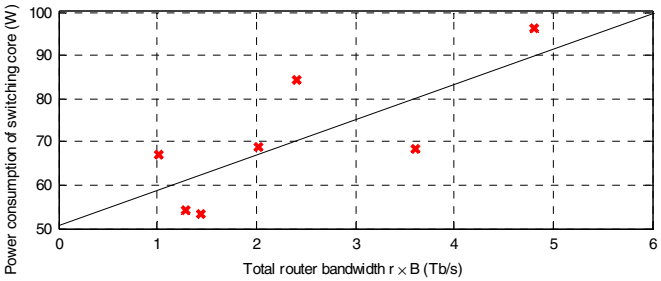

Fig. 2. Power consumption of switching cores of Table III as function of total bandwidth.

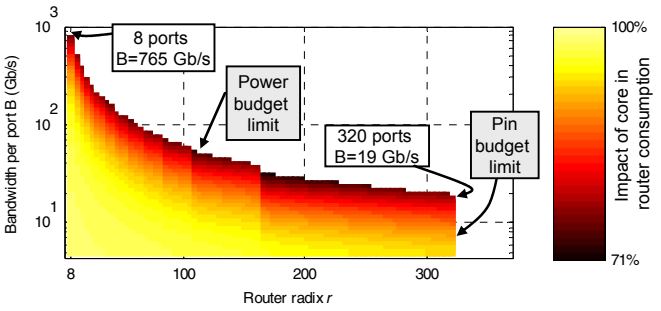

Fig. 3. Feasibility of switches with 132W and 1280 pin envelopes.

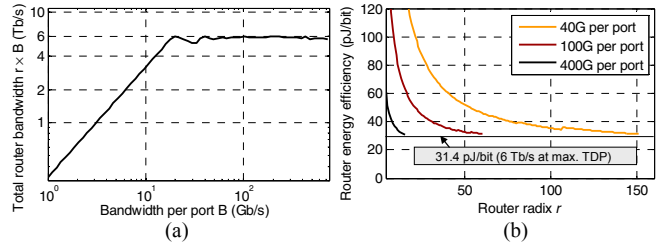

Fig. 4. a) Total bandwidth limitations as function of bitrate per port B b) Router energy efficiency as function of radix for various bitrates per port B.

TABLE II. PROPERTIES OF COMMERCIAL ROUTERS (CHIP POWER = 70% OF TOTAL POWER)

Prod.	Model	Ports	Line-rate (Gb/s)	Total BW (Tb/s)	Power (W)	Chip power (W)
Mellanox	M3601Q	32	40	1.28	85	59.5
Mellanox	SX6015	18	56	1.008	103	72.1
Mellanox	SX6025	36	56	2.016	113	79.1
Mellanox	SB7700	36	100	3.6	136	95.2
Intel	12200	36	40	1.44	85	59.5
Intel	Omni-P.	24	100	2.4	146	102.2
Intel	Omni-P.	48	100	4.8	189	132.3

TABLE III. POWER ESTIMATIONS FOR COMMERCIAL ROUTERS

Model	Chip power (W)	Chip eff. (pJ/bit)	IOBs. eff. (pJ/bit)	Core eff. (pJ/bit)	Core power (W)
M3601Q	59.5	46.48	4.25	42.23	54.06
SX6015	72.1	71.53	5.11	66.42	66.95
SX6025	79.1	39.24	5.11	34.13	68.80
SB7700	95.2	26.44	7.48	18.97	68.29
12200	59.5	41.32	4.25	37.07	53.38
Omni-P.	102.2	42.58	7.48	35.1	84.26
Omni-P.	132.3	27.56	7.48	20.09	96.42

IV. Architecture Exploration

We now leverage our models for topologies and components to perform end-to-end analysis of the HPC interconnect power consumption. We define a supercomputer as a triplet (Π_{total}, $\Pi_{compute}$ and v). Π_{total} is the maximal aggregated computing power the supercomputer can deliver (peak power) and, as already mentioned, $\Pi_{compute}$ is the maximal computing power delivered by a single compute node. We thus have $\Pi_{total} = N\ \Pi_{compute}$. The verbosity factor v indicates how the interconnect capabilities are scaled compared to computing capabilities. This definition of a supercomputer permits us to calculate the rate of node-router (NR) links ρ_{NR} as $\rho_{NR} = v\ \Pi_{compute}$.

The way the N compute nodes are interconnected is defined by another triplet (C, κ and ϕ). As already introduced in Section II, C is the concentration factor that dictates how many compute node share one entry switch, and κ describes how NR and RR differ in bitrate ($\rho_{RR}=\kappa\rho_{NR}$). ϕ is a newly introduced parameter indicating the share of RR links being optical ones [8].

Our global model for HPC interconnects is illustrated in Fig. 5. We assume the consumption of all electrical transceivers attached to routers (R of RR type, operating at rate ρ_{RR}, and C of NR type with bitrate ρ_{NR}) to be captured by the router model as described above. Note that for the cases with $\kappa \neq 1$, i.e. bitrates are imbalanced, we sweep all possibilities to split router pins available among NR and RR ports and take the least consuming option. We also substitute $C\rho_{NR} + R\rho_{RR}$ to rB in Eq. (5) or (6). On top of the router consumption, the consumption of N transceivers at rate ρ_{NR} must additionally be accounted for, because each node has one emitter and one receiver thus one transceiver. Each optical link must also be surrounded by electrical transceivers, as mentioned earlier. We therefore account for the consumption of $SR\phi$ additional optical *and* electrical transceivers.

We start the exploration by considering a Π_{total} = 20 PF system with the compute node power $\Pi_{compute}$ ranging from 0.5 to 10 TFs and maximal supported verbosity of 0.01 byte/FLOP. We initially fix the concentration factor to C=2 and κ = 1 to mimic a Cray Gemini interconnect. We finally assume that half of the links need optical transmission [8]. Results in terms of power consumption are shown in Fig. 6. We first note that no results are available for $\Pi_{compute}$ > 7.5 TFs. Above this limit the consumption of the router chip exceeds the 132W threshold.

The smallest contributors to power consumption are the NR links. Growing node compute power leads to less nodes N. This leads to concentrate the total injection traffic, which remains constant at $v\ \Pi_{total}$ = 200TB/s, in less NR links. This in turn results in growing NR link power consumptions from 75mW with $\Pi_{compute}$ = 0.5 TF to ~3W with $\Pi_{compute}$ = 7.5 TF. Since there are N=2,667 nodes in this latter case, the total consumption of the NR links is 3W·2,667 = 8 kW. Optical RR links are the second least consumers, they consume about twice as much as the NR links. At such scales, the ideal R corresponding to C=2 is in the 8…10 range. For $\Pi_{compute}$ = 7.5 TF, R=8. In this case, there are 8S RR links and 4·S *optical* RR links. As C=2 and N=2·S, we have 2·N *optical* RR links. The consumption of the optical links is, however, slightly more than twice the one of NR links (19.3 W), as one optical transceiver, of efficiency 1pJ/bit, is accounted additionally for. The largest consumers are the routers. Their individual power consumption goes up as bitrate scales with $\Pi_{compute}$ (from ~55W to ~130W), but their number decreases. As mentioned in previous section, the "static" power consumption of routers γ_{CORE} is better amortized when the router bandwidth is maximized and so is its power consumption. We note that for $\Pi_{compute}$ = 1.5 TF, the system has some resemblances with TITAN. At this point, the forecasted interconnect power consumption is 629kW. This is 7.6% of the power announced for TITAN [11]. The model thus agrees reasonably with this data-point.

Figure 7 shows the resulting global interconnect energy efficiency for C=2 but also for C=5 and C=15. Best efficiencies, within a particular C value, are always obtained with the largest compute nodes (~171 pJ/bit for C=2, ~133 pJ/bit for C=5 and ~106 pJ/bit for C=15). However, efficiencies tend to improve with larger C, both in absolute terms and for a specific $\Pi_{compute}$ value (Fig. 8a).

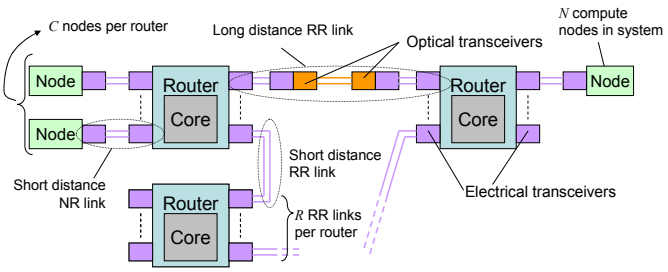

Fig. 5. Global interconnect model.

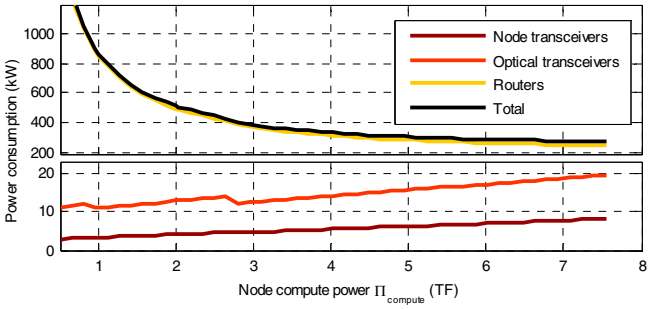

Fig. 6. Power consumptions for 20 PF interconnect (dimensioned for workload verbosity v= 0.01 byte/FLOP), for various node compute strengths.

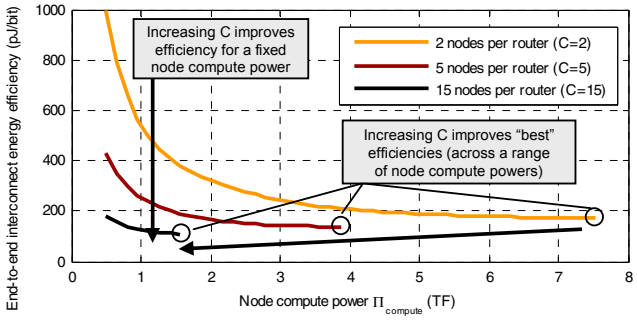

Fig. 7. Energy efficiency of 20 PF interconnects (v = 0.01 byte/FLOP), as function of node compute strengths, for various concentration factor C.

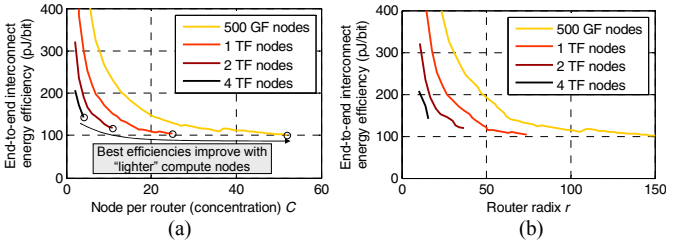

Fig. 8. Energy efficiency of 20 PF interconnects (v = 0.01 byte/FLOP) a) as function of concentration factor C b) as function of the router radix r.

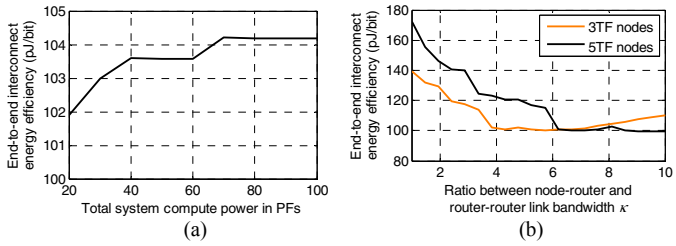

Fig. 9. a) Minimum interconnect energy efficiency as function of system scale. b) Impact of link imbalance factor κ on end-to-end interconnect energy efficiency.

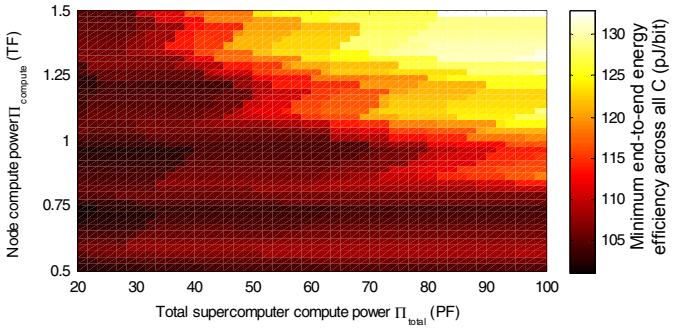

Fig. 10. Minimum interconnect energy efficiency as function of system scale Π_{total} and node compute strength.

As apparent in Table I, larger C values cause the radix to increase which permits to fully leverage the router chip power budget (up to 132W) with numerous low to medium rate ports (corresponding to lower node compute powers) instead of few high-rate ports (requiring e.g. $v\,\Pi_{compute}$ = 0.01 byte/Flop · 7.5 TF = 75GB/s = 600 Gb/s). This makes embedded electrical transceivers less consuming, offering more power budget for switching, which permits to accommodate the larger radixes. In the space explored, best efficiencies (~104 pJ/bit) have been found for C=21 and $\Pi_{compute}$ = 1.1 TF, C=27 and $\Pi_{compute}$ = 0.9 TF and for C=35 and $\Pi_{compute}$ = 0.7 TF (radixes: 62, 80 and 103 ports). In all these cases, Δ~2, i.e. each bit goes through 3 routers in average. The efficiency of routers is ~31 pJ/bit. This means that ~11 pJ/bit are consumed in the electrical and optical transceivers. The total bandwidth of the routers is 5.5 Tb/s (C=21) and 5.75 Tb/s (C=27 or 35). For C=35, this correspond to 5.75/103 = 55.8 Gb/s per port. With 103 ports, 12 pins can be allocated to each port, translating in 3 channels each carrying 18.6 Gb/s. At this speed, our electrical link power model returns 5.02 pJ/bit. Each bit using ~2 non-router included electrical transceiver and one optical transceiver, we obtain 11.02pJ/bit of "no router" consumption.

We now investigate how the scaling of the total system power Π_{total} (in FLOPS) affects efficiency. We calculated the interconnect energy efficiency for a variety of node compute powers and concentration factors. Least energy-per-bit values encountered are depicted in Fig. 9a. Somehow surprisingly, the energy efficiency is marginally affected by the system scale. However, achieving this ideal *interconnect* efficiency requires increasingly smaller compute node powers. To attain 110 pJ/bit, 1.4 TF nodes are acceptable with Π_{total} = 20 PetaFLOPS (PF) but for Π_{total} = 100 PF, compute nodes must deliver less than 0.8 TF (Fig. 10). This may go against energy efficiency objectives as such small compute nodes may show themselves bad pJ/FLOP figures. Furthermore, this obliges the full supercomputer to scale beyond 100,000 nodes, a sheer number that might raise other concerns. To reduce the number of nodes while continuing to benefit from energy efficient high radixes, imbalance between NR and RR links (κ) can be introduced. Fig. 9b shows how the best efficiency (among various C, and for constant Π_{total} = 100 PF) is affected by κ. For κ=*1,* ideal energy efficiency is above 139 and 172 pJ/bit for 3 and 5 TF nodes respectively. Bringing κ to around 6 (for 3 TF) or 7 (for 5TF) brings the energy efficiency down to ~100 pJ/bit again. Link imbalance is thus useful to decouple energy efficiency from node compute strength.

V. Discussion

The fact that the topology model covers only direct topologies is not a strong limitation. First, direct and indirect topologies are *in fine* governed by the same rules, especially in terms of router per node or link per node [9]. Furthermore, models relating the amount of resources required in any indirect topology could also be envisaged. In return, power results provided by the methodology correspond to hypothetical GMG topologies. The transition to an approaching practical topology will irremediably translate into a modification of the power figures. Numbers related here must thus be interpreted in relative terms.

Undeniably, the router model defines global results to a large extend, in particular through the two "hard" limits of 132W and 1280 pins. If power is absolutely privileged over cost, much larger chip areas, in turn allowing larger TDP and pinouts, can be envisioned.

Having the router consumption scaling with rB exclusively does not dictate specific trade-offs between bandwidth and ports: the consumption is minimized as long as the total bandwidth is maximized. Fixing for instance η_{CORE} to 1 mW/port^2 slightly changes the situation. We plan to reproduce these experiments with more sophisticated router models, typically based on finely detailed micro-architecture models [24,28].

Models and results presented here bypass most conventions in terms of radix (often a product of 2s and 3s), line-rates (in general, 10, 40, 56 or 100G) or cable formats (generally based on 4 or 8 pairs). Ignoring the latter is probably the most problematic as ad-hoc cables are unlikely to be developed. However, the methodology can also be applied while forcing some parameters to take specific values. By comparing with

"free range" configurations, cost of standard commitment can be evaluated.

Substantial power savings can be achieved by re-balancing compute power across more or less compute nodes, and distributing bandwidth across bitrate optimized links. However, such adaptations will likely impact the supercomputer overall performance, and might provide no benefits in energy-to-solution terms. Quantifying this impact is also included in our future work.

In general, the methodology presented here permits to prescreen regions of interest in the design space and seize main relationships between consumption and high-level parameters. If used in conjunction with more accurate power models, it can provide interesting support for HPC system designers in the early stages of the development.

VI. Conclusion

We proposed a methodology to predict the power consumption of any HPC interconnect, based on high-level characteristics. The methodology relies on a topology model that relates the interconnect high-level characteristics to the number and type of networking equipment required. The methodology also requires equipment power consumption models, several instances of which have been presented.

We showed how our methodology can be applied to PetaFLOP scale supercomputer designs. Results exhibit the dominance of routers in end-to-end power consumption, and the marginal role of optical links. The opportunity of using different bitrates for core or access interconnect links has also been underlined.

Acknowledgment

Authors acknowledge the financial support of the U.S. Department of Energy (DoE) National Nuclear Security Administration (NNSA) Advanced Simulation and Computing (ASC) program through contract PO1426332 with Sandia National Laboratories. Furthermore, the authors thank the CEA LETI for their support on defining the power consumption models.

References

[1] S. Rumley, M. Glick, S. D. Hammond, A. Rodrigues, K. Bergman, "Design Methodology for Optimizing Optical Interconnection Networks in High Performance Systems," *ISC High Performance,* 2015.

[2] G. Faanes, et al. "Cray cascade: a scalable HPC system based on a Dragonfly network", *SuperComputing*, 2012.

[3] M. Besta, T. Hoefler, " Slim Fly: A Cost Effective Low-Diameter Network Topology", *SuperComputing*, 2014.

[4] J. Kim, W. J. Dally, B. Towles, A. K. Gupta, "Microarchitecture of a High-Radix Router", *ISCA '05*, 2005.

[5] T. Hoefler, E. Jeannot, G. Mercier, "An Overview of Topology Mapping Algorithms and Techniques in High-Performance Computing", Emmanuel Jeannot and Julius Zilinskas. *High Performance Computing on Complex Environments*, Wiley, 2014.

[6] O. Tuncer, V. J. Leung, A. K. Coskun, "PaCMap: Topology Mapping of Unstructured Communication Patterns onto Non-contiguous Allocations", *ICS'15*, 2015.

[7] A. Daryin, A. Korzh, " Early evaluation of direct large-scale InfiniBand networks with adaptive routing", in *Journal of Supercomputing Frontiers and Innovations*, 1(3), 2015.

[8] C. Minkenberg, "HPC Networks: Challenges and the Role of Optics", *Optical Fiber communication/National Fiber Optic Engineers Conference,* 2015.

[9] G. Kathareios, C. Minkerberg, B. Prisacari, G. Rodriguez, T. Hoefler, "Cost-Effective Diameter-Two Topologies: Analysis and Evaluation", *SuperComputing*, 2015.

[10] *www.**green500**.org/*

[11] *www.**top500**.org/*

[12] S. Rumley, D. M. Calhoun, S. D. Hammond, A. F. Rodrigues, K. Bergman, " Toward Transparent Optical Networking in Exascale Computers", *ECOC Conference*, 2015.

[13] D. A. B. Miller, "Rationale and Challenges for Optical Interconnects to Electronic Chips", in *Proc. IEEE* 88(6), 2000.

[14] R. S. Tucker, "Green Optical Communications—Part I: Energy Limitations in Transport," in *IEEE Journal of Selected Topics in Quantum Electronics*, (17)2, 2011.

[15] R. S. Tucker, "Green Optical Communications—Part II: Energy Limitations in Networks," in *IEEE Journal of Selected Topics in Quantum Electronics*, (17) 2, 2011.

[16] J. B. Héroux *et al*., "Energy-Efficient 1060-nm Optical Link Operating up to 28 Gb/s," in *Journal of Lightwave Technology*, 33(4), 2015.

[17] S. Nakagawa, D. Kuchta, C. Schow, R. John, L. A. Coldren and Y. C. Chang, "1.5mW/Gbps Low Power Optical Interconnect Transmitter Exploiting High-Efficiency VCSEL and CMOS Driver," *Optical Fiber communication/National Fiber Optic Engineers Conference,* 2008.

[18] J. E. Proesel, B. G. Lee, C. W. Baks and C. L. Schow, "35-Gb/s VCSEL-Based optical link using 32-nm SOI CMOS circuits," *Optical Fiber communication/National Fiber Optic Engineers Conference,* 2013.

[19] J. E. Proesel, B. G. Lee, A. V. Rylyakov, C. W. Baks and C. L. Schow, "Ultra-low-power 10 to 28.5 Gb/s CMOS-driven VCSEL-based optical links [Invited]," in *IEEE/OSA Journal of Optical Communications and Networking*, 4(11), 2012.

[20] M. Bahadori, R. Polster, S. Rumley, Y. Thonnart, J.-L. Gonzalez-Jimenez, K. Bergman, "Energy-Bandwidth Design Exploration of Silicon Photonic Interconnects in 65nm CMOS," *IEEE Optical Interconnects Conference*, 2016.

[21] R. Polster, Y. Thonnart, G. Waltener, J. L. Gonzalez, and E. Cassan, "Efficiency Optimization of Silicon Photonic Links in 65-nm CMOS and 28-nm FDSOI Technology Nodes," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, in press, 2016.

[22] S. Scott, D. Abts, J. Kim, W. J. Dally "The BlackWidow High-Radix Clos Network", *ISCA* '06, 2006.

[23] G. Passas, M. Katevenis, D. Pnevmatikatos, "Crossbar NoCs Are Scalable Beyond 100 Nodes", IEEE TCAD, 31(4), 2012.

[24] N. Binkert, A. Davis, N. P. Jouppi, M. McLaren, N. Muralimanohar, R. Schreiber, J. H. Ahn, "The Role of Optics in Future High Radix Switch Design", *ISCA* '11, 2011.

[25] M. J. Ellsworth, Jr., R. E. Simons, "High Powered Chip Cooling— Air and Beyond," *Electron. Cooling*, 11 (3), 2005.

[26] J. G. Koomey, "Estimating total power consumption by servers in the US and the world", Technical Report, Lawrence Berkeley National Laboratory, Feb. 2007.

[27] H. Wang, L.-S. Peh and S. Malik, "Power-driven design of router microarchitectures in on-chip networks," *Annual IEEE/ACM International Symposium on Microarchitecture,* 2003.

[28] H.-S. Wang, L.-S. Peh, S. Malik, "A Power Model for Routers: Modeling Alpha 21364 and InfiniBand Routers", IEEE Hot Interconnects, 2002.