ハイパフォーマンスコンピューティング適用に向けた光インターコネクションの展望

青木 剛*

Prospect of Optical Interconnect Technology for High Performance Computing

Tsuyoshi AOKI*

* 株式会社富士通研究所 (〒 243-0197 神奈川県厚木市森の里若宮 10-1)

* Fujitsu Laboratories Ltd. (10-1 Morinosato-Wakamiya, Atsugi, Kanagawa 243-0197)

1. はじめに

近年の科学技術の進展に伴い、新規材料開発、気象予測、創薬設計、化学反応予測、および自動車や建築物の設計など、さまざまな分野において高性能計算（ハイパフォーマンスコンピューティング）が重要度を増してきている。ハイパフォーマンスコンピューティングとは巨大システムによる膨大な計算処理を指し、これにより得られる高精度予測で上記の開発コストや開発期間の削減を可能としている。上記のハードウェアの一つとして、スーパーコンピュータが用いられており、日進月歩の勢いで高性能化が進んでいる。高性能化のためにはインターコネクションの高速化が必要である。現在のスーパーコンピュータシステムでは従来の銅配線による電気インターコネクションだけ難しくなっており、光配線を用いた広帯域の光インターコネクションが導入されつつある状況である。近年のスーパーコンピュータにおける光インターコネクションについて信頼性の観点から述べる。

2. スーパーコンピュータにおける光インターコネクション

スーパーコンピュータの性能指標によるランキングの一つとして Top500¹⁾ が多く用いられている。これは LINPACK²⁾ ベンチマークプログラムにおいて連立一次方程式を解くスピードである浮動小数点演算性能 (Floating Operations Per Second, FLOPS) を基準にランキングしたものである。年二回の頻度で更新されており、最新の 2013 年 11 月時点で 10 PFLOPS 以上の高速演算が可能なスーパーコンピュータシステムは 4 つ存在する。1 位は天河二号³⁾ (NUDT, 中国) で 33.8 PFLOPS, 2 位は Titan⁴⁾ (オークリッジ国立研究所, アメリカ), 3 位は Sequoia⁵⁾ (国家核安全保障局およびローレンスリバモア国立研究所, アメリカ), 4 位は京⁶⁾ (独立行政法人理化学研究所, 日本) である。最近では、2013 年秋にアップグレードした TUBAME 2.5⁷⁾ (東京工業大学, 日本) は単精度での理論最高性能値として、17.1 PFLOPS

を達成できる性能を有している。10 PFLOPS は 1 秒間に 1 京回の演算が可能であり、これは、全人口 70 億人が 24 時間休まず、1 秒間に 1 回計算して約 17 日間行ったものに相当する驚くべき膨大な計算量⁸⁾ である。

システムでの演算高速化のためには、インターコネクションの高速化が重要な要件の一つである。近年のシステムにおいては、ノードと呼ばれるプロセッサやメモリで構成される計算単位は大規模に並列されており、ノード間での通信やストレージなどと通信するためのインターコネクションが使用されている。インターコネクションの伝送速度の高速化に伴い、従来の電気銅配線では損失増大による信号強度の減衰で伝送距離が著しく短くなってきている。それに代わり、光トランシーバと呼ばれる部品で、電気信号を光信号に変換し、光ファイバなどの伝送路で接続することで、配線での損失をほぼゼロとし、広帯域かつ長距離での接続を可能とする光インターコネクションと呼ばれる光配線技術が普及し始めている状況である。

スーパーコンピュータに既に導入されている光トランシーバの形態としては、ラック間などの長距離でアクティブ光ケーブル (Active Optical Cable, AOC) を用いたものが多い。AOC とは、インターフェースとしてケーブルの両端に電気コネクタを有しており、コネクタ部で電気信号を光信号に変換し、光ファイバを用いて光信号が伝送されるケーブル部品である。ボード側の電気コネクタはボード端に搭載されており、従来の電気ケーブルの感覚で使用できる利点がある。一部 IBM のシステムにおいてボード搭載型の光モジュールが実用化され始めている。ボード搭載型は LSI (Large Scale Integration) チップの周囲に配置することで AOC よりも電気配線長を短縮することが可能である。今後、従来よりもさらに短距離 (ラック内、ボード内) の光インターコネクション用として普及していくことが考えられる。将来的には、要求伝送速度や消費電力の観点から、従来よりも多くの接続で光インターコネクションの導入がなされていくと考えられる。

TOP500 によるスーパーコンピュータの性能トレンドで

図 1. スーパーコンピュータの実効計算性能 (RMax)

は 10 年で約 1,000 倍の速度で計算性能が向上している¹⁾。図 1 にスーパーコンピュータの実効計算性能 (RMax) を年代別にプロットしたものを示す。1 位のものは、2011 年に 10 P (ペタ) FLOPS を達成し、2019 年頃には 1 E (エクサ) FLOPS を実現する見込みとなっている。エクサスケールの実現に向けたハードウェアの課題としては、大きなものとしてシステムの大規模化に伴う電力増大、部品数増によるコスト増、さらにはそれらに伴う故障時の保守の工数増などが挙げられる。現行の構成での単純な拡張では実現困難であり、課題解決のためのさまざまな技術開発が必要となっている⁹⁾。

光インターコネクションの利用に関して言えば、今後の要求速度の高速化に伴い、伝送容量および電力の観点から、電気インターコネクションよりも今後優位になってくると考えられるが、その際の光部品導入のコストや部品の信頼性の懸念がある。前者のコストに関して言えば、部品自身のコスト削減努力は言うまでもなく必要であるが、今後スーパーコンピュータだけではなく、サーバなどでも利用されることで、物量効果により徐々に低減されていくと考えられる。後者の信頼性に関しては、システムの観点からどの程度の値が要求されていくのであろうか。現在から将来に向けてシステムから要求される光インターコネクションへの信頼性に焦点を絞って論じる。

3. スーパーコンピュータの信頼性

スーパーコンピュータは巨大なシステムであるため、膨大な数の部品によって構成されている。したがって、構成部品の故障率 FIT¹⁰⁾ (Failure In Time: $1/10^9$ 時間) が小さいものを用いても、平均故障時間間隔 MTBF¹⁰⁾ (Mean Time Between Failures: 時間) は著しく短くなる。一例として、仮に 10,000 FIT のパソコンがあるとする場合、その MTBF は 100,000 時間であり、年に表すと 11.4 年である。これは

表 1. Cray スーパーコンピュータのキャビネット数、性能、リポート頻度およびそれらの MTBF のデータ¹¹⁾

システム	Cabinet 数	性能 (Tflops)	リポート回数 (／月)	MTBF (日)
JaguarPF XT5	200	2,331	11.4	2.6
Franklin XT4	102	356	5.2	5.7
Kraken XT5	88	1,029	4.5	6.6
Jaguar XT4	84	260	6.8	4.4
Hopper XE6	68	1,289	3.8	8.0
Athena XT4	48	165	2.8	10.7
Kraken XT4	48	165	3.0	10.0
Raptor XE6	30	410	3.0	10.0
Hexagon XT4	15	51	0.6	52.1
Gaea XT6	14	260	2.6	11.7

表 2. 2009 年のシステムの故障頻度¹²⁾ および 10 ペタスケールにした場合の MTBF 見積

	故障数／月／TFLOPS	故障数／日／10 PFLOPS	10 PFLOPS での MTBF
Cray XT3/XT4	0.1~1	33~3,333	26 sec~44 min
IBM POWER 5/6	1.3	433	3.3 min
Clusters (x86-64/AMD64)	2.6~8.0	867~2,667	32 sec~1.7 min
Blue Gene L/P	0.01~0.03	3.3~100	14.4 min~7.3 hrs

11.4 年に 1 回故障する確率であることを示している。このパソコンを 10,000 台接続した並列計算システムとして稼働させる場合、システムの MTBF は僅かに約 10 時間となり、1 日約 2 回の故障確率に相当する。過去のスーパーコンピュータシステムの MTBF として表 1 に開示されている Cray のスーパーコンピュータのキャビネット数、性能、リポート頻度およびそれらの MTBF のデータ¹¹⁾を示す。性能向上につれてキャビネット数が増大し、徐々にシステムの MTBF が短くなっていく傾向がみられる。2.3 PFLOPS で MTBF が既に 2.6 日迄短縮化されており、大規模化に伴い故障はかなりの頻度で発生することがわかる。

2009 年段階でのスーパーコンピュータシステムの開示されている故障率データ¹²⁾を表 2 の左列に示す。スーパーコンピュータの 1 TFLOPS あたりの FIT 数は非常に小さいことがわかるが、ペタスケールでは数日に一回の故障が発生する確率を示している。例えば現在のスーパーコンピュータシステムの性能である 10 PFLOPS を当時のシステムのままでの拡張による大規模化で実現した場合の故障率および MTBF をデータから見積もった。それぞれの見積結果を、表 2 の中列および右列に併せて示す。TFLOPS あたりの FIT 数から単純にスケールを大きくするだけでは、10 PFLOPS での MTBF は数秒から 7 時間迄と非常に短い時間で 1 回の頻度で故障が発生する確率を示している。単純なスケールアップでは大規模システム構築は困難であることが示されている。実際の現行の 10 PFLOPS クラスのスーパーコンピュータの MTBF については、未だ正確な公表値はないが、ベンチマーク性能テストである LINPAC での計算時間や故障データから見積もられた MTBF は Sequoia で 19 時

間¹³⁾、天河2号で5時間以上¹⁴⁾、京で23時間以上¹⁵⁾と推測されている。これは、アーキテクチャでの工夫や水冷システムなどの冷却技術を駆使してジャンクション温度の低下による信頼性確保、冗長配線やネットワークの迂回経路が存在するためである。これらにより、いくつかの部品が故障しても直ぐにシステムダウンとなることはない^{16),17)}。しかしながら、計算パフォーマンスの低下や部品交換の運用、保守コストを考慮すると、システムの信頼性設計は非常に重要である。

4. 光インターコネクション部品への信頼性要求

現状のスーパーコンピュータへ光インターコネクションを導入すると仮定した場合、光部品への信頼性目標値はどの程度必要であろうか。現行のシステムの一例として、ここでは京を取り挙げて、その信頼性目標値の見積を試みた。京は約90,000のノードをもつ超巨大システムである。ノードとは並列システムにおける計算単位を指し、各ノードの構成要素はCPU、メモリ、ネットワークチップおよび電源供給ユニットなどである。京のインターコネクションはTOFU (Torus Fusion) と呼ばれる6次元メッシュトラス結合を持ち、各ノードが6軸の構成となっている¹⁶⁾。TOFUインターコネクションの接続方向はA, B+, B-, C, X+, X-, Y+, Y-, Z+, Z-で10方向からなる。したがって、1ノードあたりの配線は10 ch となる。また、TOFUインターコネクションではそれらのchで5 GB/s の速度で電気配線により双方向接続されている。

京において上記TOFU接続を光配線にする場合の信頼性を考える。約90,000ノードを稼働させるためのノードあたりのMTBF対要求FIT数の関係を図2に示す。MTBFの長時間化に伴い、要求FIT数は急激に減少していくことがわかる。例えば、24時間故障なく運用させるためには、1ノードあたり約460 FITが必要であり、72時間運用のためには約150 FITが必要であることがわかる。ノードは前述のようにCPUやメモリ、その他の電気部品を含んでいる。京の電気部品の具体的な故障率は開示されていないが、既に開示されているIBMのBlue Gene/Lでの1ノードあたりの電子部品の故障率データ¹⁸⁾を表3に示す。構成は全く同一ではないが目安として1ノードあたりで算出すると約77 FIT/ノードとなる。さらに、京は水冷の為、水冷され

ている部品(CPUなど)のFIT数が低減されており、その効果を加味して、ここでは約50 FIT/ノードと仮定した。前述のノードあたりの要求FIT数から仮定した電気部品のFIT数を差し引き、ノードあたりに使用する光インターコネクションのch数で割ると、光インターコネクションのchあたりのFIT数の見積が得られる。

前述のTOFUインターコネクションでは10 ch 存在し、1 ch あたり5 GB/sでの双方向伝送を行っている。伝送方法により多少の差異はあるが、ここでは40 Gb/sに相当すると仮定する。現行の10 Gb/s 駆動の一般的な垂直共振器面発光レーザ (Vertical Cavity Surface Emitting LASER, VCSEL) による光インターコネクションの接続を用いると、1ノードあたりの必要最低限のVCSELによる光インターコネクションの接続数は1 ch あたり4個必要であり、全てのchに配置すると、ノードあたり4個×10 ch=40個必要である。各VCSELによる光インターコネクションの接続方向は異なるが、仮にファイバシートなどで、ファイバの配線組換えを行い、1ノードに1つの光コネクタで接続される構成と仮定する。すなわち最低個数のみで全ての接続を可能とする構成とした場合を仮定する。本仮定を基に光インターコネクションのchあたりのFIT数を算出した。MTBFと光インターコネクションのchあたりのFIT数を図3に示す。京のLINPACK計算時間は約23時間であり、その際、

図2. 90,000ノードシステムでのノードあたりのFIT数とMTBFの関係

表3. IBMのBG/Lでの1ノードあたりの電子部品のFIT数データ¹⁸⁾

部 品	部品あたりの故障率 (FIT)	システムでの部品数 (個)	システムでの故障率 (FIT)	ノードあたりの故障率 (FIT)
CTL FPGA	160.0	3,024	483,840	7.3
DRAM	5.0	608,256	3,041,280	45.7
Compute + I/O ASIC	20.0	66,560	1,331,200	20.0
LINK ASIC	25.0	3,072	76,800	1.2
Clock chip	6.5	1,200	7,800	0.1
Power supply	500.0	384	192,000	2.9
Total	—	—	5,132,920	77.2

図3. システムでの使用を想定した光配線 ch あたりの FIT 数と MTBF の関係

全て故障なく駆動させるための光インターコネクションの ch あたりの要求 FIT 数は、およそ 10 と非常に小さい値であることが見積もられる。

現在の VCSEL を用いた AOC で報告されている ch あたりの故障率は 40°C で約 30 であり^{19),20)}、得られた値は現行の 1/3 と小さいことがわかる。今回の計算には多くの仮定が入っており精度は不十分ではあるが、大まかに見積もってみても光インターコネクションに関して非常に高信頼性が求められることがわかる。逆に、現行のスーパーコンピュータシステムでの大量使用が出来ない理由として、信頼性は一つの要因となっている可能性があるともいえるであろう。

5. 今後のエクサスケール実現に向けて

現在のスーパーコンピュータの規模でも、光インターコネクションへの信頼性要求値の見積りは十分高いが、今後エクサスケールのスーパーコンピュータでは、さらにその信頼性要求が高まることが予測される。ノード数はエクサスケールに向けてどのくらい増えるのであろうか。表4にTOP500における1位から5位までのシステムでのノード数およびコア数^{21)~25)}を示す。これを見る限りでも、ノード数およびコア数はシステム性能と必ずしもリンクしていない。これはさまざまなネットワークポロジやそれらのアーキテクチャ^{26),27)}、高速の演算を得意とするGPU (Graphics Processing Unit) の利用^{28),29)}、および1CPUあたりのコア数増⁹⁾などの幾つかの開発トレンドがあるためである。したがって、1EFLOPSでのノード数は単純に10PFLOPSの100倍にはならないと予測⁹⁾されている。

エクサスケールでのノード数の見積りを得るため、過去から現在までの推移からの推定を試みた。表5にTOP500での2000年から2010年までのシステム構成から得られた年平均成長率のデータ³⁰⁾を示す。CPUクロックとコアパフォー

表4. TOP500における1位から5位までのシステムでの実効演算速度、ノード数およびコア数^{21)~25)}

順位	システム	ピーク性能 (PFLOPS)	ノード数	コア数
1	天河二号	33.86	約 16,000	約 312 万
2	Titan	17.59	18,688	56 万 640
3	Sequoia	17.17	98,304	157 万 2,824
4	京	10.51	88,128	70 万 5,024
5	Mira	8.57	49,152	78 万 6,432

表5. TOP500における2000年から2010年までのシステム構成から得られた年平均成長率³⁰⁾

	LINPACK 性能	コア数	ノード数	CPU クロック	コア パフォーマンス
年平均成長率	192%	147%	133%	119%	113%

マンスは横ばいになってきており、本データからもコア数とノード数で規模を拡張していく傾向がみられる。2010年でのノード数を基にエクサスケールが実現する2019年頃でのノード数は約10万から最大約50万程度の規模と予測した。先の京の90,000ノードの接続で見積では約10であったため、エクサスケールのノード間インターコネクションを全て光で行くとノード数に反比例して、chあたり約2から9と非常に厳しい値となることが予測される。ただし、この値は現行のLINPACK計算時間23時間を基に算出した値であること、水冷による信頼性への効果の見積が正確ではないこと、また、配線冗長性やシステム迂回経路を全く加味していないことを仮定していることから、あくまで参考程度の数字であることを付記しておく。

例えば、システム側でも上述の冗長配線やエラー対策などの施策で光インターコネクション構成部品の故障率の許容量を増やすことも考えらえる。さらには、光インターコネクションの爆発的な広帯域化が進み、ノードあたりの光インターコネクションを構成する部品の数が低減する可能性も考えられる。例えば、シリコンフォトリソと呼ばれる技術が近年注目されている。これは、シリコン基板上のシリコン層を導波路に加工し、超小型の受光素子や変調器を一つのデバイスとする技術である。通信分野で用いられている波長多重技術(Wavelength Division Multiplexing, WDM)も構成によっては利用可能であり、広帯域かつ高密度の光インターコネクションが実現できる可能性がある。既に、東工大のTSUBAMEのシステムでは、シリコンフォトリソはAOC形態で用いられている³¹⁾。

システム設計と光部品技術の両者の進展で、今後のエクサスケールスーパーコンピュータでの大規模超高速光インターコネクションが実現され、近い将来のビッグデータ解析や他のさまざまな分野でのハイパフォーマンスコンピューティングの利用による高度なICT社会の到来を期待したい。

(2014.4.6- 受理)

文 献

- 1) <http://www.top500.org/>
- 2) <http://www.top500.org/project/linpack/>
- 3) J. Dongarra: "Visit to the National University for Defense Technology Changsha, China," <http://www.netlib.org/utk/people/JackDongarra/PAPERS/tianhe-2-dongarra-report.pdf>
- 4) B. Bland, et al.: "Titan: Early experience with the Cray XK6 at Oak Ridge National Laboratory," https://cug.org/proceedings/attendee_program_cug2012/includes/files/pap138-file2.pdf
- 5) B. Carnes, et al.: "Science at LLNL with IBM Blue Gene/Q," IBM J. Res. & Dev., Vol. **57**, No. 1/2, pp. 11:1–11:18, 2013
- 6) M. Yokokawa, et al.: "The K Computer Japanese Next-Generation Supercomputer Development Project," '11 Proceedings of the 17th IEEE/ACM ISLPED, pp. 371–372, 2011
- 7) <http://www.titech.ac.jp/research/stories/tsubame2013.html>
- 8) <http://jp.fujitsu.com/about/tech/k/whatis/system/>
- 9) <http://open-supercomputer.org/wp-content/uploads/2012/03/hpci-roadmap.pdf>
- 10) 日本工業規格, JIS Z 8115:2000 デイペンダビリティ (信頼性) 用語
- 11) P. Kovatch, et al.: "The Malthusian Catastrophe Is Upon Us! Are the Largest HPC Machines Ever Up?," Euro-Par 2011: Par. Processing Workshops Lecture Notes in Comput. Sci., Vol. **7156**, pp. 211–220, 2012
- 12) H. D. Simon: "Petascale Computing in the U.S.," ACTS Workshop Berkeley, California, 2006
- 13) http://news.mynavi.jp/articles/2012/12/17/sc12_sequoia/index.html
- 14) <http://wired.jp/2013/06/10/chinese-supercomputer-destroys-speed-record-and-will-get-much-faster/>
- 15) <http://news.mynavi.jp/articles/2011/07/01/kcomputer/index.html>
- 16) Y. Ajima, et al.: "Tofu: Interconnect for the K computer," FUJITSU Sci. Tech. J., Vol. **48**, No. 3, pp. 280–285, 2012
- 17) T. Noguchi, et al.: "Major Features and Outline of Supercomputer SX-8," NEC Tech. Journal, Vol. **58**, No. 4, pp. 7–10, 2005
- 18) P. Coteus, et al.: "Packaging the Blue Gene/L supercomputer," IBM J. Res. & Dev., Vol. **49**, No. 2/3, pp. 213–248, 2005
- 19) K. Takaki, et al.: "1060nm VCSEL for inter-chip optical interconnection," Proc. of SPIE, Vol. **7952**, pp. 795204-1-795204-4, 2011
- 20) Finisar® APPLICATION NOTE, "10Gb Oxide Isolated VCSEL Reliability Report," <http://www.finisar.com/sites/default/files/pdf/6iJfN610Gb%20Oxide%20Isolated%20VCSEL%20Reliability%20Report.pdf>
- 21) <http://wired.jp/2013/06/10/chinese-supercomputer-destroys-speed-record-and-will-get-much-faster/>
- 22) 宮崎博行, 他: "スーパーコンピュータ「京」の概要," FUJITSU, Vol. **63**, No. 3, pp. 237–246, 2012
- 23) http://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu2/006/shiryo/_icsFiles/afieldfile/2013/05/27/1335363_09.pdf
- 24) <http://eetimes.jp/ee/articles/1306/20/news044.html>
- 25) <https://www.alcf.anl.gov/user-guides/mira-cetus-vesta>
- 26) D. J. Kerbyson, et al.: "A performance comparison of current HPC systems: Blue Gene/Q, Cray XE6 and InfiniBand systems," Future Generat. Comput. Syst., Vol. **30**, pp. 291–304, 2014
- 27) <http://news.mynavi.jp/articles/2007/01/01/supercomputer/002.html>
- 28) J. E. Stone, et al.: "GPU-accelerated molecular modeling coming of age," J. Mol. Graph. Model., Vol. **29**, pp. 116–125, 2010
- 29) M. Macedonia: "The GPU Enters Computing's Mainstream," Computer, Vol. **36**, No. 10, pp. 106–108, 2003
- 30) 西川 岳: "今後の HPC 技術に関する研究開発の方向性について," http://www.mext.go.jp/b_menu/shingi/chousa/shinkou/028/shiryo/_icsFiles/afieldfile/2012/06/13/1322027_2.pdf
- 31) S. Matsuoka: "TSUBAME 2.0 から TSUBAME 2.5, 3.0, 更にはエクサへの進化 TSUBAME2.0 to 2.5, Towards 3.0 and Exascale," GPU technology conference, July 2013, <http://on-demand.gputechconf.com/gtc/2013/jp/sessions/1002.pdf>

著者紹介

青木 剛 (あおき つよし)

1998年 京都大学・工学。2000年 同大大学院工学系研究科修士課程了。同年 (株) 富士通研究所入社。現在, 基盤技術研究所 フォトニックデバイス研究部においてサーバ, スパコン向け光インターコネクションの実装に関する研究開発に従事。2005年 東京大学博士 (工学) 取得。