دانشکده مهندسی کامپیوتر

عنوان

ارائه درس مدارهای واسط (پروتکل InfiniBand)

دانشجو

محمدپارسا بشری ۴۰۰۱۰۴۸۱۲

استاد

دکتر امین فصحتی

پاییز ۱۴۰۳

فهرست مطالب

منبع اصلی این ارائه White Paper ارائه شده توسط خود شرکت Nvidia است. منابع دیگر نیز به عنوان منابع تکمیلی (مثلا برای مباحث RDMA) استفاده شده است:

1. White paper: https://network.nvidia.com/pdf/whitepapers/IB_Intro_WP_190.pdf
2. InfiniBand Trade Association website: https://www.infinibandta.org/
3. Wickipedia: https://en.wikipedia.org/wiki/InfiniBand
4. https://www.fs.com/blog/infiniband-what-exactly-is-it-7714.html
5. https://community.fs.com/encyclopedia/remote-direct-memory-access-rdma.html
6. https://www.fibermall.com/blog/how-to-choose-between-infiniband-and-roce.htm
7. LinkedIn post by Pawan Sharma: https://www.linkedin.com/pulse/infiniband-vs-fiber-channel-ethernet-pawan-sharma-9qhtc/

همچنین چهار ویدیوی زیر را از یوتیوب مشاهده کردم:

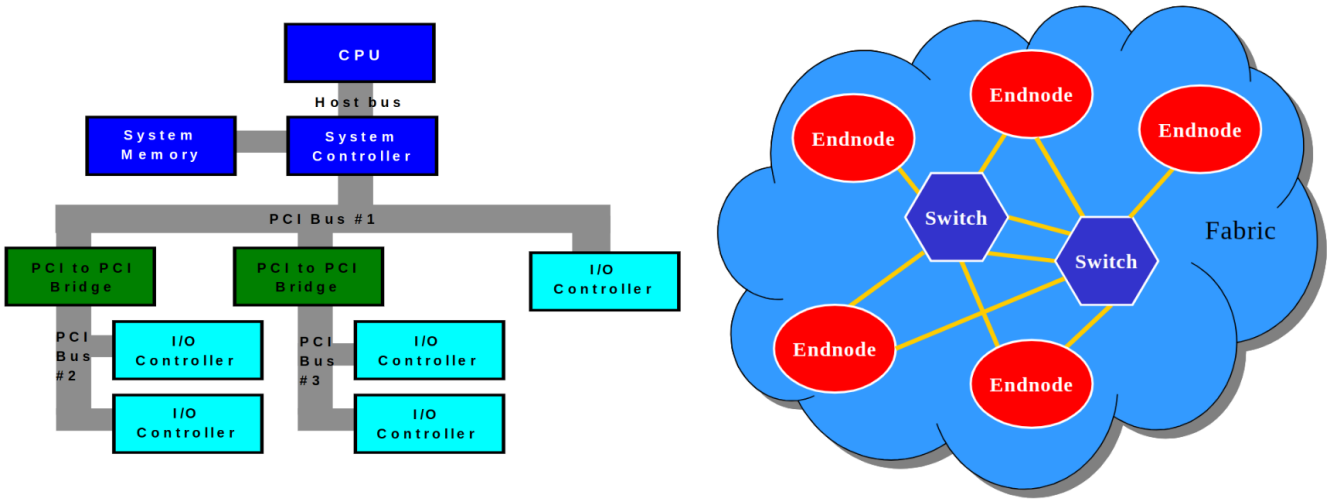
1. https://www.youtube.com/watch?v=OW7fbBt-wVE
2. https://www.youtube.com/watch?v=cowASe-dc7o
3. https://www.youtube.com/watch?v=eGoP2wPoaEM
4. https://www.youtube.com/watch?v=xXXrX1CcuBw

۱ مقدمه

اینفینی‌بند (InfiniBand) یک فناوری شبکه با کارایی بالا و تأخیر کم است که عمدتاً در محیط‌های HPC، خوشه‌های AI و مراکز داده ابری مورد استفاده قرار می‌گیرد. این فناوری ارتباط مؤثر بین گره‌های محاسباتی را فراهم کرده و پهنای باند بالا و مقیاس‌پذیری زیادی را ارائه می‌دهد. یکی از ویژگی‌های کلیدی اینفینی‌بند کاهش بار پردازنده مرکزی از طریق پشتیبانی از RDMA است که امکان انتقال داده مستقیم بین حافظه‌های سیستم‌های مختلف را فراهم می‌کند. نکته قابل توجه درباره اینفینی‌بند این است که این پروتکل هم برای ارتباطات درون کامپیوتری و هم بین کامپیوتری استفاده می‌شود و می‌تواند به نرخ داده تا ۴۰۰ گیگابیت بر ثانیه دست یابد.
اینفینی‌بند معماری منحصربه‌فردی دارد که شامل اجزای مختلفی مانند آداپتورهای کانال میزبان (HCAs)، سوئیچ‌ها و روترها است. به دلیل قابلیت‌های پیشرفته اینفینی‌بند، از آن در سیستم‌های پردازش با کارایی بالا و ابررایانه‌ها استفاده گسترده‌ای می‌شود؛ به طوریکه در لیست TOP500 از بین ۱۰۰ ابرکامپیوتر برتر، ۶۳ تا از آنها از اینفینی‌بند استفاده می‌کنند.
یکی از ویژگی‌های مهم اینفینی‌بند، مدل ارتباطی آن است که بر اساس لایه‌های مختلف پروتکلی پیاده‌سازی شده است. لایه فیزیکی مشخصات کابل‌ها و سرعت انتقال داده را تعریف می‌کند، در حالی که لایه لینک مسئول کنترل جریان و تشخیص خطاها است. در لایه شبکه، مسیریابی بسته‌ها مدیریت می‌شود و لایه انتقال گزینه‌های ارتباطی مطمئن و نامطمئن را ارائه می‌دهد.
با توجه به عملکرد بالا و تأخیر بسیار کم، اینفینی‌بند یک جایگزین قدرتمند برای اترنت (Ethernet) در محیط‌های محاسباتی پیشرفته محسوب می‌شود. مقایسه اینفینی‌بند با فناوری‌های مشابه نشان می‌دهد که این پروتکل در سیستم‌هایی که نیاز به انتقال سریع داده و پردازش بی‌درنگ دارند، مانند رایانش ابری، هوش مصنوعی و تحلیل داده‌های کلان، عملکرد بهتری ارائه می‌دهد.

۲ معماری InfiniBand

اینفینی‌بند یک پروتکل مبتنی بر Switched Fabric است. شکل ۱ معماری سنتی گذرگاه مشترک را با معماری Switched Fabric مقایسه می‌کند.

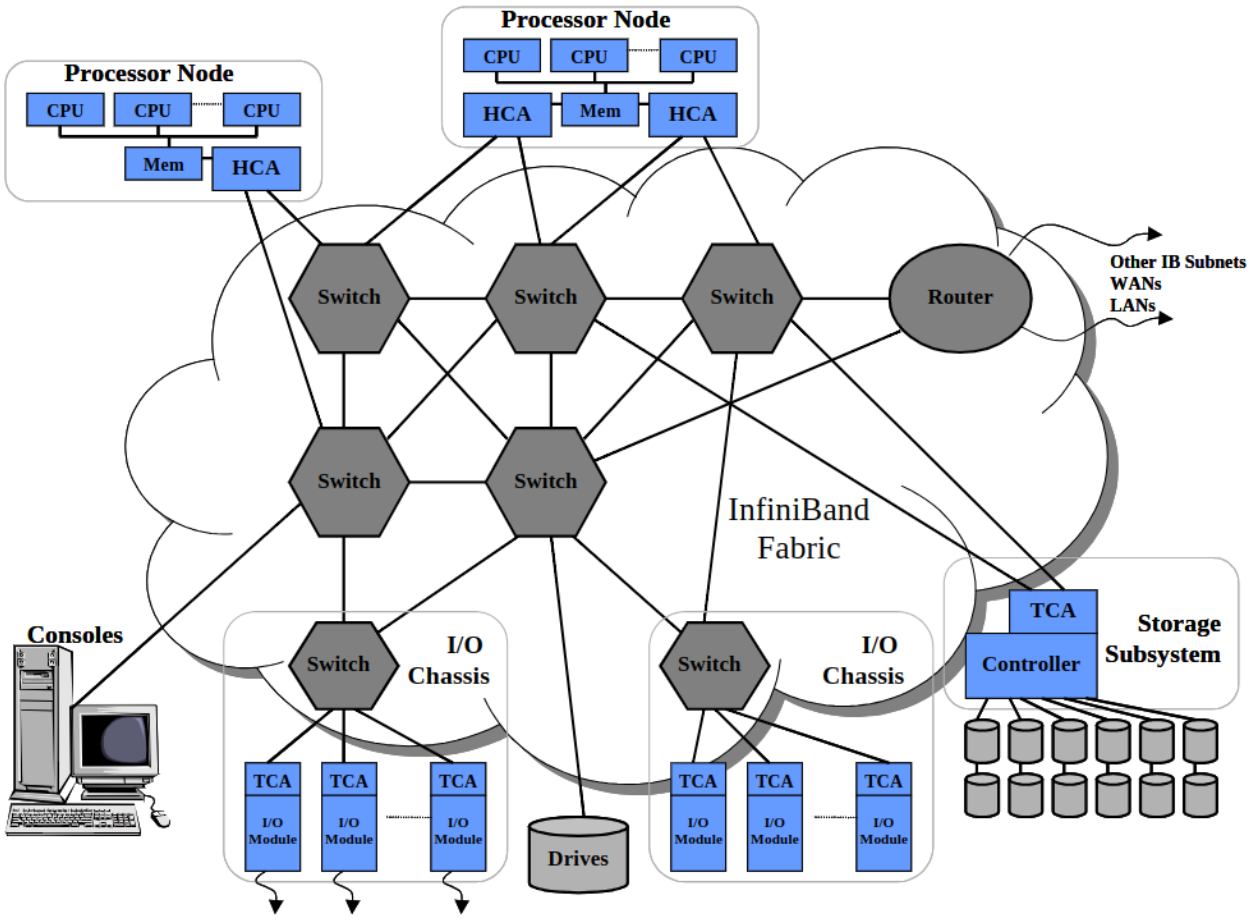


شکل ۱: مقایسه معماری گذرگاه مشترک (سمت چپ) و Switched Fabric (سمت راست).

در معماری Switched Fabric تعدادی Endnode داریم که می‌توانند پردازنده‌ها، حافظه‌ها و یا کنترلرهای I/O باشند. این معماری که به نوعی point-to-point محسوب می‌شود، به وسیله سوییچ‌ها این Endnodeها را به یکدیگر متصل می‌کنند. به مجموعه‌ای از این Endnodeها که به وسیله سوییچ به یکدیگر متصل شده‌اند و از یک آدرس‌دهی داخلی مشترک استفاده می‌کنند، یک Subnet می‌گویند. برای ارتباط بین Subnetها از روترها استفاده می‌شود که در لایه شبکه کار می‌کنند (لایه‌های پروتکل اینفینی‌بند در قسمت‌های بعدی شرح داده می‌شود).

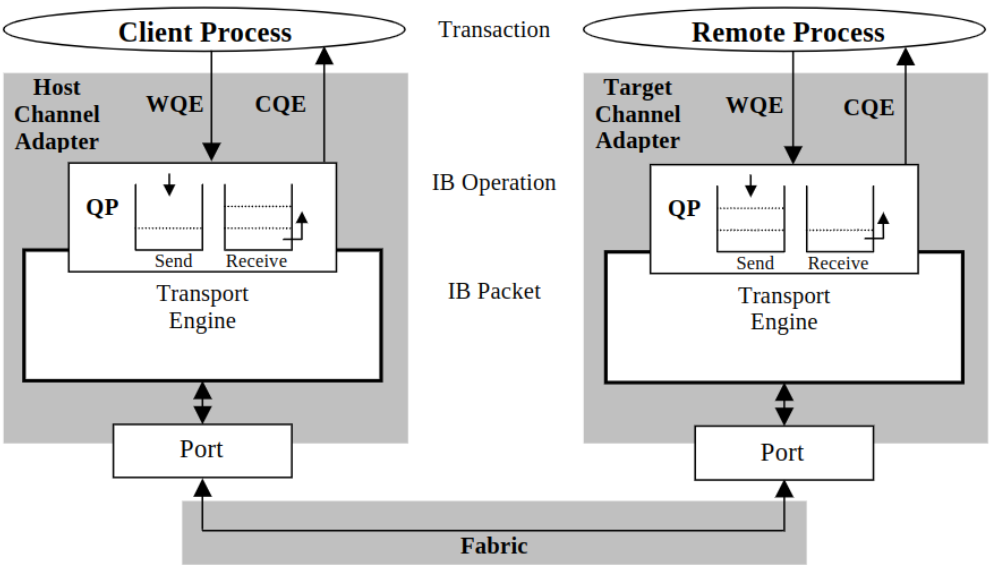
شکل ۲ معماری پروتکل اینفینی‌بند را نشان می‌دهد. در این معماری چهار کامپوننت اصلی وجود دارد:

- **آداپتور کانال میزبان (HCA)**: یک رابط سخت‌افزاری است که وظیفه اتصال سرورها به شبکه اینفینی‌بند را بر عهده دارد. HCA از RDMA پشتیبانی کرده و ارتباطات با تأخیر کم و پهنای باند بالا را امکان‌پذیر می‌کند.
- **آداپتور کانال هدف (TCA)**: مشابه HCA است اما معمولاً در دستگاه‌های ذخیره‌سازی و سخت‌افزارهای جانبی به کار می‌رود. این آداپتور امکان انتقال داده مستقیم بین حافظه دستگاه‌های مختلف را فراهم می‌کند.
- **سوئیچ (Switch)**: یک دستگاه شبکه‌ای است که ارتباط بین چندین گره را مدیریت کرده و بسته‌های داده را به صورت هوشمند بین مسیرهای مختلف هدایت می‌کند.
- **روتر (Router)**: روترها برای اتصال چندین Subnet اینفینی‌بند به یکدیگر استفاده می‌شوند. آن‌ها بسته‌های داده را بین شبکه‌های مختلف ارسال کرده و ارتباطات در مقیاس بزرگ‌تر را امکان‌پذیر می‌کنند.



شکل ۲: معماری پروتکل اینفینی‌بند.

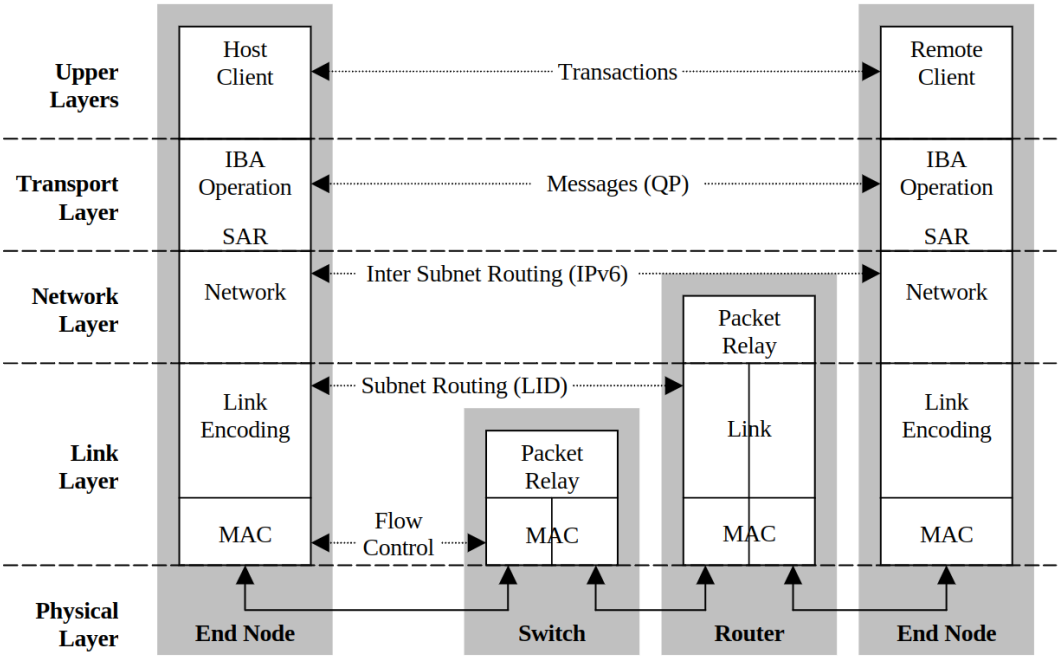
معماری داخلی HCA و TCA در شکل ۳ نمایش داده شده است. هر کدام از این کامپوننت‌ها شامل یک جفت صف (QP) است که یکی از آنها برای ارسال و دیگری برای دریافت پکت‌ها استفاده می‌شوند. اپلیکیشن در حال اجرا در پردازه‌ی لایه بالاتر، با استفاده از Work Queue Entry یا WQE بسته‌های ارسالی را به صف ارسال می‌فرستد و صف دریافت بسته‌های دریافت‌شده را توسط Completion Queue Entry یا CQE به پردازه مربوطه ارسال می‌کند. با استفاده از این روش، این کامپوننت‌ها می‌توانند وضعیت بسته‌های ارسالی و دریافتی را مانیتور کنند و از انتقال صحیح بسته‌ها مطمئن شوند.



شکل ۳: ساختار داخلی HCA و TCA در اینفینی‌بند.

۳ ساختار لایه‌ای پروتکل InfiniBand

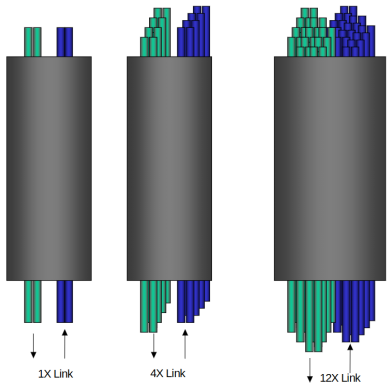
پروتکل اینفینی‌بند از یک ساختار لایه‌ای مشابه OSI بهره می‌برد. این ساختار لایه‌ای در شکل ۴ نشان داده شده است. این ساختار شامل ۵ لایه اصلی است که هر کدام مستقل از دیگری وظیفه مشخصی بر عهده دارد. در ادامه این بخش، هر کدام از لایه‌ها را مستقلا بررسی می‌کنیم.



شکل ۴: ساختار لایه‌ای پروتکل اینفینی‌بند.

۱.۳ لایه فیزیکی

اینفینی‌بند سه سرعت متفاوت برای انتقال در لایه فیزیکی تعریف می‌کند: 1X، 4X و 12X. هر لینک از چهار سیم تشکیل شده که یک ارتباط full duplex تفاضلی با نرخ ۲/۵ گیگابیت بر ثانیه را تشکیل می‌دهند (یک جفت سیم برای ارسال و یک جفت سیم برای دریافت). شکل ۵ شمایی از این سه مدل لینک را نشان می‌دهد و شکل ۶ آنها را مقایسه می‌کند.



شکل ۵: لینک‌های فیزیکی مختلف در پروتکل اینفینی‌بند.

InfiniBand Link	Signal Count	Signalling Rate	Data Rate	Fully Duplexed Data Rate
1X	4	2.5 Gb/s	2.0 Gb/s	4.0 Gb/s
4X	16	10 Gb/s	8 Gb/s	16.0 Gb/s
12X	48	30 Gb/s	24 Gb/s	48.0 Gb/s

شکل ۶: مقایسه سه لینک مختلف در لایه فیزیکی اینفینی‌بند.

۲.۳ لایه لینک

لایه لینک به همراه لایه انتقال، قلب تپنده پروتکل اینفینی‌بند را تشکیل می‌دهد. این لایه مسئولیت ارتباط بین Endnodeها در یک Subnet را بر عهده دارد و مسائلی مانند انواع پکت، کنترل جریان داده، تشخیص خطا و ... را پیاده‌سازی می‌کند.

بسته‌ها: در لایه لینک، دو نوع بسته وجود دارد: (۱) بسته‌های مدیریتی و (۲) بسته‌های داده. بسته‌های مدیریتی برای کانفیگ کردن و نگهداری لینک استفاده می‌شوند و به طور مثال اطلاعات دیوایس‌ها توسط این نوع بسته‌ها منتقل می‌شود. بسته‌های داده می‌توانند تا ۴ کیلوبایت از داده لایه‌ی بالاتر را در خود جای دهند.

سوییچ‌ها: داخل هر Subnet، مسئله فوروارد کردن پکت‌ها توسط سوییچ‌ها در لایه لینک هندل می‌شود. هر دیوایس داخل Subnet یک Local ID (LID) ۱۶ بیتی دارد که توسط Subnet Manager به آن داده می‌شود. هر پکت دارای یک Local Route Header یا به اختصار LRH است که شامل LID مقصد است و مسیریابی داخل هر Subnet توسط این فیلد صورت می‌گیرد.

کنترل جریان داده: اینفینی‌بند از یک مکانیزم Credit Based برای کنترل جریان داده بین دو Endnode استفاده می‌کند. هر Endnode مقداری داده‌ای که می‌تواند بدون پر شدن بافرش دریافت کند را توسط بسته‌های لایه لینک مشخصی advertise می‌کند و فرستنده بر اساس آن بسته و مقدار داخلش، نرخ ارسال بسته به آن Endnode را تنظیم می‌کند.

یکپارچگی داده: در هر بسته دو CRC برای تشخیص خطا وجود دارد. فیلد Variant CRC یا به اختصار VCRC یک فیلد ۱۶ بیتی است که CRC تمام فیلدهای داخل پکت را محاسبه می‌کند. این نوع CRC با عبور از هر سوییچ یا روتر دوباره محاسبه می‌شود و به همین دلیل Variant نامیده شده است. فیلد Invariant CRC یا به اختصار ICRC یک فیلد ۳۲ بیتی است و فقط فیلدهایی از پکت را کاور می‌کند که با عبور از روترها یا سوییچ‌ها تغییر نمی‌کنند.

۳.۳ لایه شبکه

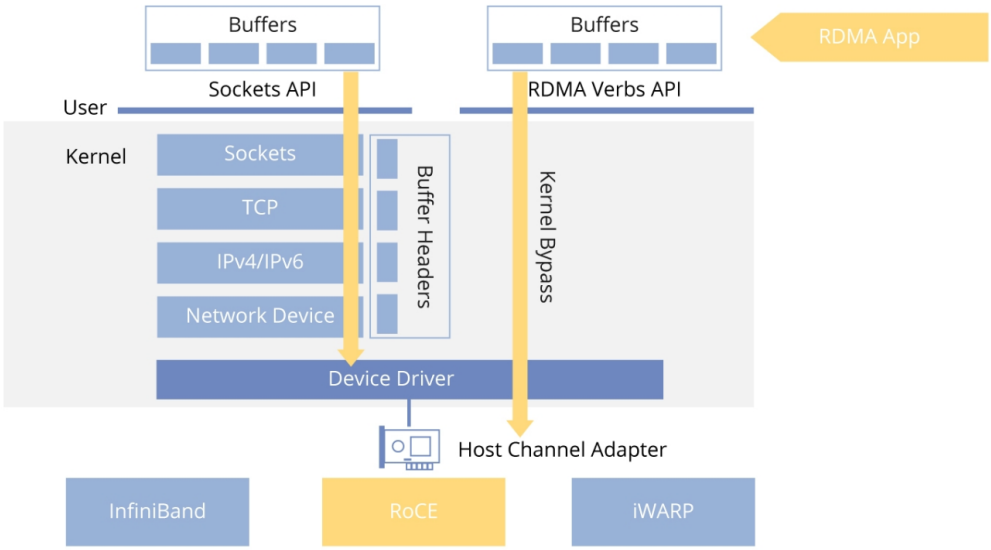
لایه شبکه مسئول مسیریابی بسته‌ها از یک Subnet به Subnet دیگر است (در داخل یک Subnet، لایه شبکه موردنیاز نیست). بسته‌هایی که بین Subnetها ارسال می‌شوند، شامل یک Global Route Header (GRH) هستند. GRH شامل آدرس IPv6 ۱۲۸ بیتی برای مبدأ و مقصد بسته است. این بسته‌ها بر اساس Global Unique ID (GUID) ۶۴ بیتی هر دستگاه، از طریق یک روتر بین Subnetها ارسال می‌شوند. روتر LRH را با LID مناسب در داخل هر Subnet اصلاح می‌کند. بنابراین، آخرین روتر در مسیر، LID موجود در LRH را با LID مربوط به پورت مقصد جایگزین می‌کند. در داخل لایه شبکه، بسته‌های InfiniBand هنگامی که در یک Subnet واحد استفاده شوند (که سناریوی رایجی برای شبکه‌های حوزه سیستمی InfiniBand است)، نیازی به اطلاعات و سربار هدر لایه شبکه ندارند.

۴.۳ لایه انتقال

لایه انتقال مسئول تحویل بسته‌ها به ترتیب، تقسیم‌بندی، چندبخشی‌سازی کانال و ارائه سرویس‌های انتقال (اتصال قابل‌اطمینان، دیتاگرام قابل‌اطمینان، اتصال غیرقابل‌اطمینان، دیتاگرام غیرقابل‌اطمینان، دیتاگرام خام) است. این لایه همچنین وظیفه تقسیم‌بندی داده‌های تراکنشی هنگام ارسال و بازسازی آن‌ها هنگام دریافت را بر عهده دارد. بر اساس Maximum Transfer Unit (MTU) مسیر، لایه انتقال داده‌ها را به بسته‌هایی با اندازه مناسب تقسیم می‌کند. گیرنده بسته‌ها را بر اساس Base Trans-port Header (BTH) که شامل جفت صف مقصد و شماره توالی بسته است، دوباره سرهم‌بندی می‌کند. گیرنده دریافت بسته‌ها را تأیید می‌کند و فرستنده این تأییدیه‌ها را دریافت کرده و صف تکمیل را با وضعیت عملیات به‌روزرسانی می‌کند. یک نکته قابل توجه این است که دراینفینی‌بند تمامی این عملکردها به صورت سخت‌افزاری پیاده‌سازی شده‌اند که منجر به بهبود قابل توجهی در کارایی و سرعت سیستم می‌شود.

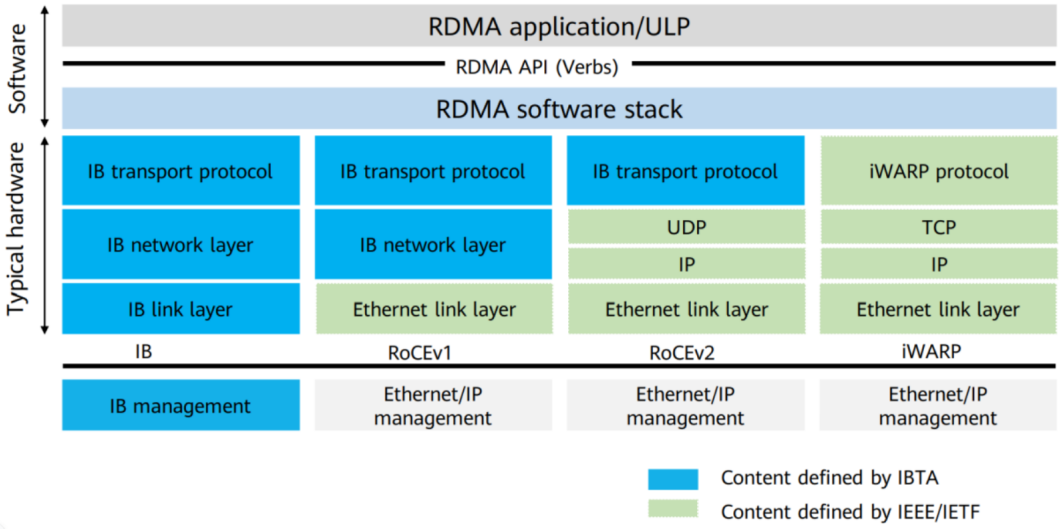
۴ روش RDMA

RDMA یا Remote Direct Memory Access یک فناوری پیشرفته برای انتقال داده است که امکان دسترسی مستقیم به حافظه یک سیستم از طریق شبکه، بدون نیاز به درگیر کردن پردازنده مرکزی، حافظه نهان یا سیستم‌عامل را فراهم می‌کند. این تکنیک موجب کاهش تأخیر، افزایش پهنای باند و بهبود کارایی سیستم‌های ارتباطی در شبکه‌های با کارایی بالا می‌شود. شکل ۷ حالت سنتی را با RDMA مقایسه می‌کند. در این شکل می‌توان دید که با دور زدن kernel می‌توان از context switch جلوگیری کرد و کارایی سیستم را به طرز چشم‌گیری بهبود بخشید.



شکل ۷: تکنیک RDMA.

در روش‌های سنتی انتقال داده، پردازنده مرکزی باید پردازش‌های متعددی از جمله کپی داده بین حافظه‌های مختلف، مدیریت وقفه‌ها و هماهنگی با سیستم‌عامل را انجام دهد که این امر منجر به افزایش سربار پردازشی و کاهش عملکرد کلی سیستم می‌شود. در مقابل، RDMA با حذف این سربار و انتقال مستقیم داده بین حافظه‌های دو سیستم، تأخیر را به حداقل رسانده و بهره‌وری شبکه را افزایش می‌دهد.
همانطور که قبل‌تر نیز گفته شد، RDMA در سه پروتکل InfiniBand، RoCE (RDMA over Converged Ethernet) و iWARP پیاده‌سازی می‌شود. شکل ۸ این روش‌ها را از لحاظ لایه‌های پروتکل مربوطه، مقایسه می‌کند.

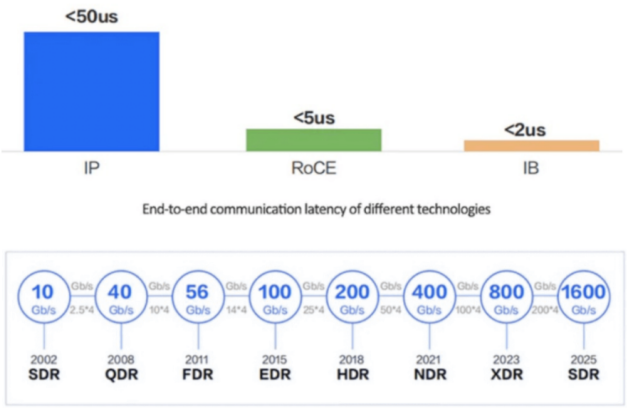


شکل ۸: مقایسه روش‌های مختلف پیاده‌سازی تکنیک RDMA.

۵ مقایسه با پروتکل‌های مشابه

در آخر به طور خلاصه به مقایسه اینفینی‌بند با Ethernet و Fibre Channel که دو پروتکل پرکاربرد دیگر در مراکز داده و پردازش ابری هستند می‌پردازیم. شکل ۹ این مقایسه را انجام می‌دهد. در جدول سمت چپ، یک مقایسه کلی با معیارهای مختلف بین این سه پروتکل انجام شده است. در شکل بالا سمت راست، مقایسه‌ای بین تاخیر end to end در سه پروتکل Ethernet، RoCE و InfiniBand نمایش داده شده است که بیانگر تاخیر بسیار کم اینفینی‌بند (به دلیل استفاده از RDMA) در مقابل دو پروتکل دیگر است. در شکل پایین سمت راست نرخ ارسال اینفینی‌بند طی سال‌های گذشته نشان داده شده است.

Feature	InfiniBand	Fibre Channel	Ethernet
Primary Use	High-performance computing	Storage area networks (SAN)	General-purpose networking
Data Rates	Up to 200 Gbps (or more)	Up to 32 Gbps	Up to 400 Gbps
Latency	Extremely low	Low to moderate	Moderate
Topology	Switch-based, scalable fabric	Fabric-oriented, dedicated SAN	Various (switched, star, etc.)
Scalability	Highly scalable	Scalable in SAN environments	Scalable, depending on architecture
Cost	Generally high	Moderate to high	Generally lower
Common Applications	HPC, data-intensive tasks	Data storage and backup	LAN, cloud, enterprise networking



شکل ۹: مقایسه اینفینی‌بند با Ethernet و Fibre Channel.