100 GigaBit Ethernet

1 Ethernet

A Ethernet é um conjunto de normas e padrões de rede que define regras numa LAN (Local Internet Network) para a transmissão de dados, implementando o algorítmo CSMA/CD (Carrier Sense Multiple Access with Collision Detection) para acesso a dados com detecção de colisão e o MAC (Medium Access control) para controle de acesso ao meio.

Esse protocolo é atualmente padronizado pelo IEEE 802.3, um grupo de estudo pertencente ao IEEE (Institute of Eletrical and Electronics Engineers), cuja a responsabilidade é estudar e padronizar esse modelo de rede, tal qual atua na camada física e de enlace de dados no modelo OSI (Open Systems Interconnection). Os padrões são especificados por velocidade, ou seja, para cada velocidade há uma normalização. Dentro da camada física do Modelo OSI, a ethernet define padrões de cabeamento, dispositivos (switches e patch panels), faixas de envio de dados e estruturas para que a velocidade desejada seja atingida. Já na camada de enlace, é usado um controlador de link lógico para destinar os dados de forma mais eficiente e também o MAC, que define frames de dados e garante que cada dispositivo conectado a rede tenha um endereço único, evitando o envio e processamento desnecessário de informações. Para interligar essas duas camadas foi desenvolvido o reconciliador e o MII (Media Independent Interface). Nesse âmbito, a 100 Gigabit, ou 100GE, é um conjunto de normas e tecnologias de rede para transmissão de dados numa velocidade de 100 Gb/s (IEEE Computer Society (2018)).

1.1 Camada física

Nesse padrão, inicialmente são determinadas as especificações da camada física (PHY - Physical Layer Device) para a transmissão desses dados, tal qual é dividida em subcamadas, são elas: Physical Coding Sublayer (PCS), Forward Error Correction (FEC), Physical Medium Attachment (PMA), Physical Medium Dependent (PMD) e

o Medium Dependent Interface (MDI).

1.1.1 Physical Coding Sublayer

A primeira subcamada física (PCS) provê o serviço de codificação/decodificação dos dados em blocos de 66bit (64B/66B), é responsável por distribuir os dados em diferentes faixas, compensação de diferença de taxas entre o reconciliador e o PMA, determinar quando uma conexão foi estabelecida informando então ao gerenciador quando o dispositivo está pronto para uso.

1.1.2 Foward Error Correction

Já na segunda subcamada física o FEC (Foward Error Correction) age com o objetivo de evitar a perda de dados através da redundância no envio de bits, onde ele faz a mesma adicionando bits ao streaming de dados pelo algorítimo Reed-Salomon, sendo então nomeado como RS-FEC (Reed Solomon Forward Error Correction). Em cada especificação o RS-FEC trabalha de uma forma e, em sua implementação na 100GE, é necessário exatamente quatro faixas de envio e outras quatro para recebimento, sendo indispensável o mapeamento 10:4 quando trabalha com o PMA possuindo 10 faixas, pois tal PMA opera com 10 faixas para envio e outras 10 para recebimento.

A terceira subcamada, o PMA (Physical Meidium Attachment), fornece o serviço de intermediação entre um PMA e um cliente, podendo esse cliente ser um PCS, FEC ou outro próprio PMA. Entre esses serviços têm-se a adaptação dos sinais das faixas dos PCS para o número de faixas físicas ou abstratas do cliente, ou seja, ele pode receber 10 faixas de stream de dados e transformá-lá em 4 faixas de stream de dados. Também provê especificações de tempo para transmissão dos dados entre as faixas assim como gerenciamento dos mesmos. O PMA faz o direcionamento de bits de dados para que todos os bits de uma stream vão e voltem pela mesma faixa. Ainda na terceira camada, quando há a comunicação entre dois PMAs, pode-se usar especificação elétrica de módulos plugáveis com dez faixas a 10.3125 GBd e também de módulos plugáveis e pontos combinados com quatro faixas a 25.78125 GBd.

A quarta subcamada (PMD) provê o serviço de intermédio entre o PMA e o MDI

controlando o envio e recebimento dos dados entre os mesmos, traduzindo o código recebido do PMA de streamings de bit para stramings elétricas ou de streamings de bits para streamings de sinais óticos e o contrário também, onde o PMA trabalha com bits e o MDI com sinais elétricos e/ou óticos. Também na implementação do PMD é decidido qual modo de comunicação/conexão usar, exemplo: Fibra ótica em Single-Mode, MultiMode ou também cabos de cobre.

Relacionado ao PMD, tem-se ainda o MDI (Medium Dependent Interface), que é a interface de comunicação entre o dispositivo PMD e o Medium, podendo o Medium ser entendido como meio de comunicação (fibra ótica, cabo de cobre, backplane). Essa interface pode ser compreendida de outro modo como o receptor e/ou transmissor acoplado ao dispositivo PMD, e varia conforme a normativa.

Já na camada de enlace, tem-se também as divisões de especificações e como principais entidades há o LLC (Logical Link Control), o MAC (Media Access Control) e também o MAC Control, que na implementação da 100GE não é necessário.

Entre as entidades, inicialmente há o MAC, que provê o serviço de transferência de dados entre MACs, onde sua semântica de transferência é constituída de: endereço de destino (que pode ser um MAC ou um grupo), endereço de origem, unidade de serviço de dados MAC e sequência de checagem de frame. Tais semânticas trabalham através de frames e pacotes sendo os frames encapsulados em pacotes pelo MAC e cada elemento é especificado conforme a tabela abaixo:

O primeiro elemento (preâmbulo), ajuda na sincronização do PLS com o tempo do pacote e serve para avisar que um frame está a caminho. O SFD é a sequência de dados fixada (10101011) que antecede o frame, ou seja, depois dela o receptor saberá que será os bits do frame. Os campos de endereço possuem 48 bits cada, e o endereço de destino pode ser um MAC unico, um grupo ou todos os endereços da LAN. O campo de Tamanho indica o número de bytes dentro do próximo campo (Dados Cliente MAC).

No campo de dados do cliente MAC, há os dados a serem transmitidos e bits de dados são adicionados ao campo para que o frame não seja eliminado no futuro como

		Quantidade de Bytes	Campo
Pacote		7 Bytes	Preâmbulo
		1 Byte	SDF
	Frame	6 Bytes	Endereço de Destino
		6 Bytes	Endereço de Origem
		2 Bytes	Tamanho
		46 a 1500 Bytes	Dados Cliente MAC (PayLoad)
		4 Bytes	Sequência de checagem de frame

Formato de Frame e Pacote Ethernet

um frame com quantidade de dados insuficiente, que é de 48 bytes. A sequência de checagem de frame (FCS) é utilizada para validação do frame e é gerada a partir de dos dados do mesmo para que haja detecção de erro no recebimento, ou seja, se o calculo da sequência no recebimento for diferente do FCS recebido, significa que o frame está errado.

Depois de encapsulado, o frame é enviado e na recepção é considerado inválido quando: seu tamanho é incondizente com o especificado no elemento de tamanho; se o frame não possuir a quantidade de bits múltipla de 8, pois deve ser uma cadeia de bytes; ou o FCS calculado não coincidir com o valor FEC recebido.

O MAC Control com CSMA/CD não se faz necessário na 100GE pois essa funcionalidade usa o algoritimo CSMA/CD (Carrier Sense Multiple Access with Collision Detection). Tal algorimo não é util na 100GE visto que ela opera semente em modo *full duplex*, logo não risco de colisão de dados.

Ainda na camada de enlace, porém acima do MAC, tem-se o LLC (Logical Link Controller) que facilita, através de mecanismos de multiplexação e demultiplexação, o trânsito e coexistência de vários pacotes num meio de rede com vários pontos. Isso é possível pois ele guarda o endereço de cada MAC dentro da rede e faz todos se enxergarem como um, ou seja, enquanto o MAC guarda a informação dos dados e dispositivos para mostrar a origem e destino do pacote, o LLC mostra o melhor caminho

a ser percorrido para esse pacote chegar ao objetivo.

Esses conceitos tecnológicos (PHY, MAC e LLD) se referem as duas primeiras camadas físicas do modelo OSI e para interligar as duas o 802.3 também padroniza o reconciliador (RS). Opcionalmente o 802.3 também padroniza as MII (Media Independent Interface), que provê a interconexão lógica entre o MAC e o PHY, atuando então embaixo do RS. O MII foi desenvolvido para que a camada de enlace de dados e o meio físico trabalhem de forma independente e é especificado na 100GE como CGMII.

Em suma, o RS converte a stream de dados dada pelo MAC para dados (sinais) paralelos do CGMII e também o mapeamento dos sinais providos pelo CGMII para as primitivas do MAC, já CGMII é o facilitador de transmissão e recebimento de dados entre o RS e o PHY.

Todas essas definições são padronizadas pela IEEE para a 100GE e vários fatores foram essenciais para o alcance de tal velocidade, isso fica claro ao comparar-lo com outros padrões como 10GE, 25GE e 400GE, sendo eles conjuntos de normas para a velocidade, respectivamente, de 10 Gb/s, 25 Gb/s e 400 Gb/s, todos eles definidos pelo grupo 802.3.

	10GE	25GE	100GE	400GE
Bloco de dados no RS (bits)	32	32	64	64
Faixas	1	1	4 ou 10	16
Velocidade por faixa (Gb/s)	10	25	25 para 4 faixas ou 10 para 10 faixas	25

Especificações de Normas 802.3

O primeiro dado se refere aos blocos de bits transmitidos através do RS, a qual se observa um aumento para o dobro do tamanho, 32 para 64 bits. A importância desse item é visto quando calcula-se a velocidade de transmissão com 10 faixas transmitindo a 156,25 Mhz: $10(faixas) \times 64(bits) \times 156,25 = 100(Gb/s)$.

Na segunda têm-se a quantidade de faixas e a velocidade por faixa. Inicialmente, em 2010, a 100GE foi padronizada com 10 faixas operando a 10 Gb/s por segundo, logo após, em 2014, a 802.3 iniciou uma força tarefa para alcançar a velocidade de 25 Gb/s de transmissão numa única faixa, tal objetivo foi atingido em 2016 quando foi aprovado tal padrão. A partir dai também foi normalizado a 25GE com uma faixa 25 Gb/s, 100GE com 4 faixas a 25 Gb/s, 200GE com 8 faixas a 25 Gb/s e a 400GE com 16 faixas a 25 Gb/s.

O conjunto de evolução de vários elementos como cabeamentos óticos (OM3, OM4 e OM5), cabos coaxiais, capacidade de processamento dos hardwares e aumento da demanda de dados a serem transmitidos foram responsáveis pelo avanço da ethernet e foi elencado dois principais, onde observa-se grande impacto dos mesmos no crescimento da ethnert e anda mais estudos estão sendo feitos para que velocidades de 1,2 Tb/s e 800 Gb/s sejam alcançadas.

Além da ethernet, há padrões de rede de alta performace como Enoc, Infiniband e Omni Path, que visam a transmissão de dados tanto para armazenamento quanto para processamento.

2 Enoc

A Rede em Chip Expansível (Enoc) é uma rede sugerida por Ivan Luiz Pedroso (2018) para interação de Sistemas num Chip (SoCs), que permite comunicação de elementos de processamento de um chip, porém esse diálogo pode se dar tanto de elementos num chip (Intra-Chip) quanto com elementos em outro chip (Inter-Chip).

Na camada física, essa rede é composta por Elementos de Processamento (PE), Ligações metálicas, buffers e reteadores, todos eles dentro de um chip. Tais membros são dispostos numa malha bidimencional onde os PEs possuem buffers para armazenar suas mensagens e esses PEs são ligados a um roteador, ou seja, há um roteador para cada PE e os roteadores são também ligados a outros quatro roteadores a sua volta. Um desses roteadores é ligado a um hub sem fio e o mesmo faz comunicação com outro hub sem fio em outro chip.

Na camada de enlace de dados, ela trabalha com roteadores, hubs e pacotes divididos em flits de 32 bits. Os pacotes são divididos e reconstruidos dentro dos PEs e enviados através dos roteadores, que possuem comunicação em bararmento full duplex. Quando o destinatário for outro chip, o flit é encaminhado ao hub sem fio através dos roteadores, que envia o mesmo para o hub do chip de destino. O pacote é dividido em 4 bytes para endereço de destino e origem, 4 a 1500 bytes para os dados a serem transmitidos (PayLoad), por fim um flit repetindo o ultimo flit do PayLoad para indicar o fim do pacote.

A Enoc é sugerida para ser expansível e reconfigurável, sendo que isso se dá através de sinais que o hub sem fio envia para informar sua presença e quando esse sinal é detectado, troca-se informações sobre seus PEs e essas informações são armazenadas dentro de cada hub, permitindo assim a expasividade sem necessidade de conhecimento prévio.

A InfiniBand (IB) é uma rede padronizada pela InfinBand Trade Association destinada para computação de alta performace, provendo um fácil meio para transporte de mensagens direto de uma aplicação a outra aplicação, storage ou sistema operacional. Enquanto a ethernet foca na transmissão de bits de dados numa rede, a IB visa criar um canal direto de comunicação, numa rede, entre elementos de uma aplicação sem necessidade de intervensao do sistema operacional (Paul Grun - InfiniBand Trade Association (2010)).

Na camada física, essa especificação é composta de Adaptador de Canal do Host (Host Channel Adapter (HCA)), Adaptador de Canal Alvo (Target Channel Adapter (TCA)), Switches, Roteadores, Cabos e Conectores.

O HCA fica num dispositivo ou computador e fornece controle e conexão para transmissão de dados com outros dispositivos, podendo ser esse segundo um HPC, TCA ou Switch. Em outras palavras, o HCA é o dispositivo físico nas pontas para o canal virtual criado entre dois pontos. O TCA promove as mesmas funcionalidades do HCA, porem de maneira mais simples, pois é feito para dispositivos com subsistemas especializados. Tal dispositivo foi, em suma, substituido pelo HCA pois este atende todas as demandas do TCA.

O switche é semelhante aos usados em outras redes, que é um disposivo para mutiplexação de pacotes, sendo diferenciado na maneira que é usado o mesmo na implementação da camada de enlace. Já os roteadores são utilizados na segmentaçõa de uma IB, ou seja, se há uma IB muito larga, ela pode ser divididas em subredes conectadas por roteadores IB.

Ainda na camada física, há os cabos e conectores, sendo os conectores o meio ao qual um sinal ótico é enviado na origem ou recebido no dispositivo de destino, ou seja, facilitam a passagem de bits elétricos para o tipo de sinal do meio. Os cabos são o meio ao qual a informação trafega, podendo ser ele uma fibra ótica, um backplane ou cabo de cobre. Todos eles trabalhando no sistema de velocidade IB.

A IB padroniza suas velocidades e faixas, trabalhando atualmente com 1, 4, 8 ou 12 faixas e velocidades de 2.5, 5, 10, 14.06 e 25.78 Gb/s. Tal combinação pode transmitir, por exemplo, 312 Gb/s se forem usados 12 faixas enviando 25.78 Gb/s.

Na camada de enlace essa rede implementa a técnica Flow Control, que consiste numa coleção de providências tomadas que um receptor não seja sobrecarregado por um dispositivo que envia numa velocidade maior. Ela é feita através de uma confirmação que o receptor envia informando ao controlador que mais pacotes podem ser recebidos.

Há vários formatos de pacotes na IB, e será descrito dois pela semplicidade de estrutura. O primeiro é o Pacote Local e é composto por 8 Bytes no cabeçalho, que contém informações como destino e origem, 0 a 4096 Bytes para PayLoad e 6 Bytes para cubrir os pacotes caso necessário. O outro tipo de pacote é o Global e sua diferença do pacote Local é o tamanho do cabeçalho, que é composto por 40 Bytes. O pacote Local é usado quando precisa-se transportar uma informação na mesma subnet e o Global de uma subnet a outra.

Tal envio é realizada por nós entre dispositivos na rede, chamados de QR (Queue Pair). Os QRs são construídos em cima de canais virtuais traçados na rede, onde um tradutor de endereço - implementado na camada de transporte - traduz o endereço virtual para o caminho físico. Ne camada de transporte, a IB oferece um tradutor de endereço virtuais para físicos e também implementações de mensagens/protocolos para que a comicação com dispositivos e storages se de forma direta e/ou facilitada.