Network Transports & Congestion Control — One-Pager

Quick reference for transport protocols (TCP/UDP family, RDMA, AI/HPC, storage, scale-up) and congestion-control algorithms used at each layer. Updated 2026.

CLASSIC IP TRANSPORTS (LAYER 4)

Protocol	Owner / Std	Reliable	Ordered	Multipath	Encryption	Key trait	Used by / for
TCP	IETF	Yes	Yes	No	External (TLS)	Byte-stream, AIMD CC, HoL blocking	Web, SSH, SMTP — universal
UDP	IETF	No	No	No	External (DTLS)	Connectionless, low overhead	DNS, DHCP, VoIP, gaming, QUIC base
SCTP	IETF	Yes	Per-stream	Multi-home failover	External (DTLS)	Multi-streaming + multi-homing	Telecom: SS7/SIGTRAN, Diameter, 5G N2
DCCP	IETF	No	No	No	No	Unreliable + congestion control	Mostly research/abandoned
QUIC	IETF (Google origin)	Yes	Per-stream	Connection migration	Built-in (TLS 1.3)	0/1-RTT setup, user-space	HTTP/3 — Google, Cloudflare, Meta, Apple
MPTCP	IETF	Yes	Yes	Yes (subflows)	External	TCP across multiple paths	Apple Siri/iOS, Samsung, Linux
UDP-Lite	IETF	No	No	No	External	Partial checksum	Loss-tolerant codecs

RDMA TRANSPORTS

Protocol	Owner / Std	Substrate	Lossless required	Multipath	Encryption	Key trait	Used by / for
InfiniBand	NVIDIA / IBTA	IB fabric	Yes (credit FC)	RD mode only	Optional	Native RDMA, sub-µs latency	DGX SuperPOD, TOP500 HPC, Meta RSC
RoCE v1	IBTA (open)	Ethernet L2	Yes (PFC)	Limited	Optional	IB transport over Ethernet, non-routable	Same-subnet RDMA
RoCE v2	IBTA (open)	UDP/IP	Yes (PFC)	Limited (ECMP)	Optional	IB transport over UDP, routable	Azure, Meta, Tencent, ByteDance, Baidu
iWARP	IETF (open)	TCP/IP	No	Via TCP	Optional	RDMA over TCP, no PFC needed	Intel E810, Chelsio (niche)
IB Verbs modes	IBTA	—	—	—	—	RC, RD, UC, UD transport modes	RC dominant in production

AI / HYPERSCALER CUSTOM TRANSPORTS (the new generation)

Protocol	Owner	Substrate	Lossless required	Multipath	Encryption	Key trait	Used by / for
MRC (Multipath Reliable Conn.)	OpenAI + AMD/MS/NV/Broadcom/Intel — OCP	Ethernet/IP	No	Yes (packet spray)	Built-in	Evolution of RoCE v2; µs failover; verbs-compat	OpenAI training, Microsoft Fairwater, Oracle Abilene
Falcon	Google — OCP	Ethernet/IP	No	Yes (PLB)	Built-in (PSP/IPSec)	HW transport, multi-ULP (RDMA + NVMe)	Google Cloud, Intel E2100 IPU
SRD (Scalable Reliable Datagram)	AWS	Ethernet/IP	No	Yes (packet spray)	Built-in	Out-of-order delivery, hw-offloaded, libfabric API	AWS EFA — EC2 P5/Trn1/Trn2/HPC
UET (Ultra Ethernet Transport)	UEC consortium (open)	Ethernet/IP	No	Yes (packet spray)	Built-in	Open standard; ~75% from HPE Slingshot; libfabric 2.0	Industry target — 1M+ endpoint scale
Pony Express	Google (legacy)	Ethernet/IP	No	Limited	Optional	SW-only predecessor to Falcon, ran in Snap microkernel	Older Google datacenter (superseded)
eRDMA	Alibaba	VPC/Ethernet	No	Yes	Optional	RDMA for cloud tenants	Alibaba Cloud ECS
HPCC	Alibaba	(CC layer, not full transport)	—	—	—	Precise CC using in-network telemetry (INT)	Alibaba RDMA datacenters

SCALE-UP & SPECIALIZED INTERCONNECTS

Protocol	Owner	Domain	Key trait	Used by / for
NVLink / NVSwitch / NVLink Fabric	NVIDIA (proprietary)	Scale-up GPU	Up to 1.8 TB/s per GPU; sub-µs	DGX, GB200 NVL72, HGX
UALink	AMD / Broadcom / Cisco / Google / HPE / Intel / Meta / MS — open	Scale-up GPU	Open NVLink alternative; v1.0 in 2025	Future open AI servers
SUE (Scale-Up Ethernet)	Broadcom	Scale-up GPU	Simpler than UET; ≤1.6 Tbps, ~100ns device latency	Broadcom AI silicon
ICI (Inter-Chip Interconnect)	Google	Scale-up TPU	Native TPU pod fabric	Google TPU v4/v5p/Trillium pods
Slingshot / Portals 4	HPE (Cray)	HPC scale-out	Adaptive routing; UET 1.0 lineage (~75%)	Frontier, El Capitan, Aurora, leadership HPC
OmniPath	Cornelis Networks (ex-Intel)	HPC scale-out	InfiniBand-style fabric	Some HPC sites
RDS (Reliable Datagram Sockets)	Oracle	Cluster IPC	Reliable datagrams over IB/RoCE/TCP	Oracle RAC interconnect
TIPC	Ericsson / Linux	Cluster IPC	Topology-aware cluster messaging	Telecom clusters

STORAGE TRANSPORTS

Protocol	Owner	Substrate	Key trait	Used by / for
Fibre Channel (FC)	T11 (open)	Dedicated SAN	Lossless block storage, deterministic	Cisco MDS, Brocade, NetApp, Dell, IBM, HPE
FCoE	T11 (open)	Ethernet	FC encapsulated in Ethernet (declining)	Legacy converged DC
iSCSI	IETF (open)	TCP/IP	SCSI over TCP, universal	Every storage vendor
iSER	IETF (open)	RDMA (RoCE/IB/iWARP)	iSCSI accelerated by RDMA	High-perf storage
NVMe-oF	NVMe Consortium	TCP / RDMA / FC	NVMe commands over fabric — near-local NVMe perf	Pure, NetApp, VAST, Weka, Dell PowerStore
SMB Direct	Microsoft	RDMA	SMB3 + RDMA (DDP)	Windows Server, Azure Files

CONGESTION-CONTROL ALGORITHMS

Algorithm	Family	Signal	Key trait	Where used
Reno / NewReno	TCP	Loss	Classic AIMD; baseline	Legacy TCP everywhere
CUBIC	TCP	Loss	Cubic growth function — default in Linux/Windows	Most internet TCP today
Vegas / Westwood	TCP	Delay / bw-est	Delay-based; low queueing	Niche; research
Compound TCP	TCP	Loss + delay	Hybrid	Older Windows
BBR v1 / v2 / v3	TCP / QUIC	Bandwidth + RTT	Model-based; fills bottleneck without filling buffers	Google services, YouTube, QUIC, Linux kernel
DCTCP	DC TCP	ECN marking	Proportional reaction to ECN; small queues	Microsoft, Linux DC TCP stacks
DCQCN	RoCE v2	ECN + PFC	Default RoCE v2 CC; rate-based	Azure, Meta, Tencent — most RoCE clusters
TIMELY	RDMA	Delay (RTT gradient)	Delay-based, CPU-light	Google early RDMA
Swift	RDMA / Falcon	Delay (NIC RTT)	Decomposes host vs fabric latency; basis for Falcon	Google Falcon
HPCC	RDMA	In-band telemetry (INT)	Precise rate using switch INT data	Alibaba
PowerTCP	DC TCP / RDMA	Power = bw × queue	Combines bandwidth and queue depth signals	Research / select DCs
MRC CC	MRC	Multipath telemetry	Programmable CC + microsecond rerouting	OpenAI/MS Fairwater/Oracle Abilene
Falcon CC (Swift+CSIG+Carousel)	Falcon	Delay + congestion sig.	HW per-flow shaping, multipath PLB	Google + Intel E2100
SRD CC	SRD	Path-level feedback	Avoids overloaded paths; <10ms RTO	AWS EFA
UET CC	UET	Sender + receiver based	Two-sided CC for packet-sprayed environment	Ultra Ethernet 1.0
Homa	Receiver-driven	Priorities + grants	Eliminates HoL via priorities; message-oriented	Stanford research, influential
NDP / Trim	Switch-assisted	Header trimming	Switch trims payload on congestion; no whole-packet drop	Cambridge research
ExpressPass	Credit-based	Receiver credits	Receiver paces with credit packets	Research
EQDS	Edge-queued	Edge-based shaping	Pushes queues to edges, not core	Cambridge/UCL research

Mental model: (1) Classic IP transports cover the internet. (2) RDMA family (IB, RoCE v2, iWARP) covers traditional HPC/AI but needs lossless fabric (PFC) and doesn't multipath well. (3) Each hyperscaler built a custom multipath transport because RoCE v2 doesn't scale to 100K+ GPUs: Google→Falcon, AWS→SRD, OpenAI/MS/NV/AMD→MRC, Alibaba→eRDMA. (4) UET is the open-standard convergence target; expect MRC/Falcon ideas to fold in over time. (5) Scale-up (NVLink / UALink / SUE / ICI) is intra-server and disjoint from scale-out transports. (6) Congestion control matters as much as the transport: most modern AI fabrics combine packet spraying + delay-based CC + ECN/INT signals + microsecond failover.