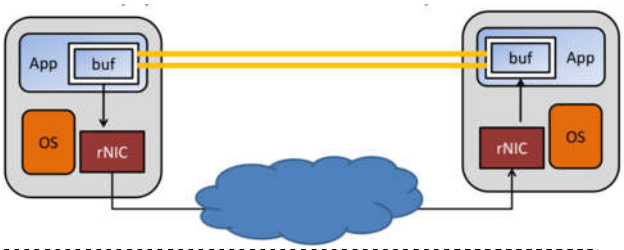
МОДУЛЬ
на компьютерном рынке России с 1991 года

ул. Сосинская 43, стр.1,3Б, Москва, 109316
+7-495-956-3499
http://www.mdl.ru
modul@col.ru

Обыкновенное чудо OmniPath. (*Первый взгляд на результаты qperf*)

https://github.com/linux-rdma/qperf инструмент тестирования производительности RDMA и IP (https://linux.die.net/man/1/qperf). Осуществляет замеры обмена между двумя узлами, выдавая результаты значений для полосы пропускания, латентности и занятости ЦПУ. Основу для преимуществ высокоскоростного обмена с низкой латентностью создаёт технология RDMA (Удаленного доступа к ОЗУ напрямую), иллюстрируемая схемой справа. Она делает возможным прямой обмен между буферами данных приложений через сетевые адаптеры минуя ЦПУ. Данная технология нашла своё применение в высокопроизводительных вычислениях, где доминирующей платформой фактически стал InfiniBand. На базе последнего позднее появилась реализация RoCE (RDMA over Converged Ethernet), основанная на промежуточном уровне IPoIB. Аналогично протоколам TCP/ UDP в обмене IP (с обратной связью и без неё), RDMA может работать с надёжными (reliable) и не надёжными (unrelaiable) соединениями: RC/ UC и датаграммами: RD/ UD. В приводимой справа таблице приводятся перечни поддерживаемых каждым из этих методов операций (глаголов, verbs) и допустимый максимальный размер сообщения для RoCE.



Операция	UD	UC	RC	RD
Send (в т. ч. напрямую)	X	X	X	X
Receive	X	X	X	X
RDMA Write (в т. ч. напрямую)		X	X	X
RDMA Read			X	X
Fetch и Add/ Cmp и Swap (атомарные)			X	X
Макс. размер сообщения	MTU	1ГБ	1ГБ	1ГБ

FDR IB	ud	rc_rdma read/ write	uc_rdma write	tcp	udp
Латентность, us	6.01			14.4	12.8
send/recv, GB/sec	5.49/ 5.48	6.13/ 6.12	6.01/ 5.98	2.78	5.44/ 3.86

Таблица слева демонстрирует типичные значения тестирования FDR IB (результаты предоставлены Ю.Шкандыбиным, МФТИ). *Характерным является существенный прирост производительности и снижение латентности при использовании RDMA.*

Рассмотрим теперь начавшую поступать на рынок с 2016г технологию Omni-Path (Intel OPA, http://www.mdl.ru/Solutions/Put.htm?Nme=CX600#OmniPath). Одним из её существенных преимуществ является возможность реализации внутри самого ЦПУ, что уменьшает, как минимум, на два хопа путь сообщения, давая преимущества в латентности. Другое преимущество, большее число портов в одном ASIC, позволяет строить большие сетевые среды меньшим числом коммутаторов и также приводит к снижению латентности.

Посмотрим теперь на результаты qperf (таблица справа, предоставлена Ю.Шкандыбиным, МФТИ). Тестирование проводилось в другой среде, поэтому сравнение со значениями из предыдущей таблицы некорректно. Нас интересует лишь сопоставление значений внутри таблицы.

OmniPath	ud	rc_rdma read/ write	uc_rdma write	tcp	udp
Латентность, us	11.9			12.3	12.8
send/recv, GB/sec	2.06/ 2.06	4.25/ 3.70	4.38/ 4.35	3.32	4.97/ 4.97

Итак, переход от обмена данными с применением RDMA к традиционному обмену TCP/ UDP IP, по крайней мере, *не приводит к падению производительности* для соответствующих показателей обмена с подтверждением и без него. Что это означает? Это говорит нам о том, что производитель провёл существенную работу по оптимизации обмена на уровне микрокодов (firmware) и, условно говоря, выполнил за программиста практически всё что нужно. Теперь вам нет нужды заботиться о выборе глаголов при общении со своей сетевой средой. Всю работу с оптимизацией обмена выполнит для вас само оборудование!

Например, у вас есть унаследованное из прошлого приложение, которое уже и непонятно как работает, а писавшие его программисты или далече, или не помнят как его создавали. Но это приложение активно применяется и вам хотелось бы улучшить его производительность в сетевой среде. Нет проблем! Ставим OmniPath и всё в шоколаде!

В настоящее время мы с коллегами приступаем к тестированию данной гипотезы на прикладных задачах, следите на http://sddc.mdl.ru