Суперкомпьютерные дни в России 2024

международная научная конференция

23 - 24 сентября

Анализ применения технологии GPU-aware MPI для сети Ангара: первые результаты

Тимур Исмагилов[1,2], Анатолий Мукосей[2], Владислав Галигеров[1,3], Юрий Гришичкин[1],

Феликс Смирнов[3], Владимир Стегайлов[1,3] и Алексей Тимофеев[1,3]

[1] Объединенный институт высоких температур РАН

[2] ВЭИ — филиал РФЯЦ–ВНИИТФ

[3] Национальный исследовательский университет «Высшая школа экономики»

RSF | Russian Science Foundation

Outline

- Angara vs InfiniBand FDR
- Desmos supercomputer
- Benchmarks: OSU and rocHPL
- ROCm clang toolchain and Score-P tracing
- Angara API in UCX for GPU-aware MPI
- Conclusions

General scope of this study

Interconnects

GPU-aware MPI communication

Analysis of parallel programs runtime

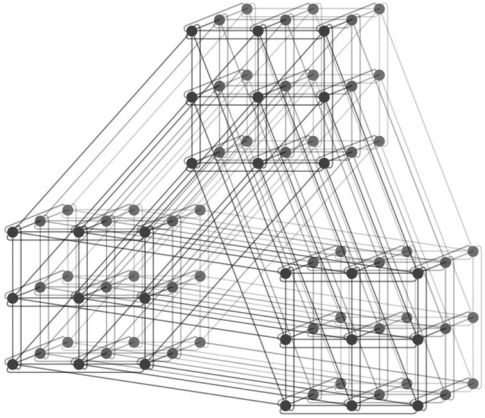
Interconnect performance vs application performance

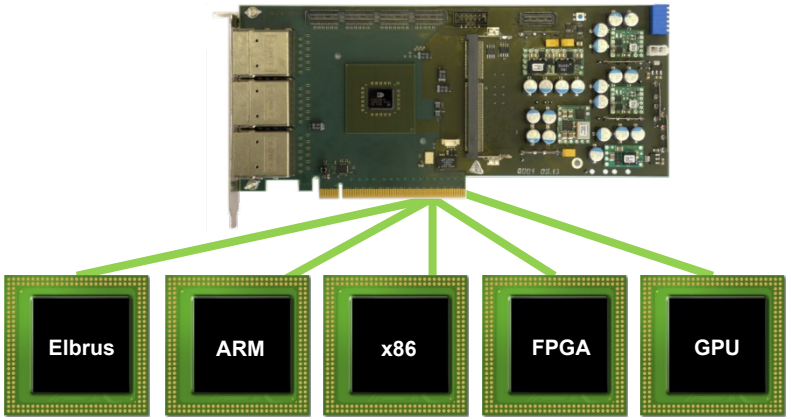
ANGARA INTERCONNECT

Angara interconnect

Key features

- Network topology: 1D..4D-torus
- ASIC-based network card
- Up to 8 communication channels
- Remote direct memory access (RDMA)
- Multi-core CPU support
- Adaptive packet transfer
- MPI ping-pong latency: 0,85 µs
- Single hop latency: 129 ns
- Scaling: up to 32K nodes
- Power consumption: up to 20 W
- Data transmission bandwidth: 5-75 Gbit/s
- Various physical transmission media



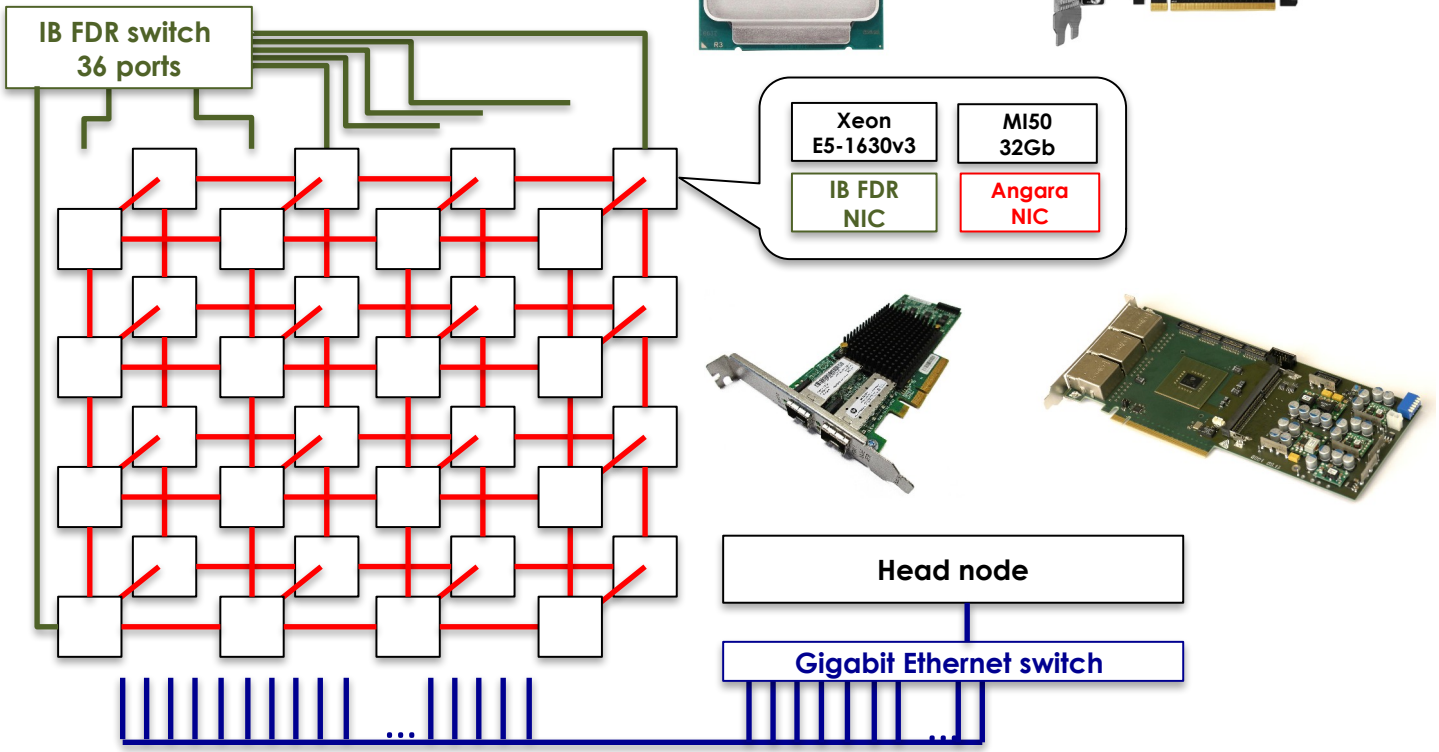


Angara interconnect

Interconnect	Mellanox IB FDR 4x	Angara	Cray Aries	Mellanox IB EDR 4x	Intel OmniPath
Year	2011	2013	2012	2015	2015
TOP500	13	–	5	20	6
Topology	fat tree / kD-torus	4D-torus	dragonfly	fat tree / kD-torus	fat tree
Bandwidth, Gbit/s	56	75	42	112	100
MPI latency, μs	1	0.85	1.3	0.92	0.9-1.0
Single hop latency, ns	– / 250	129	100	n/a	n/a

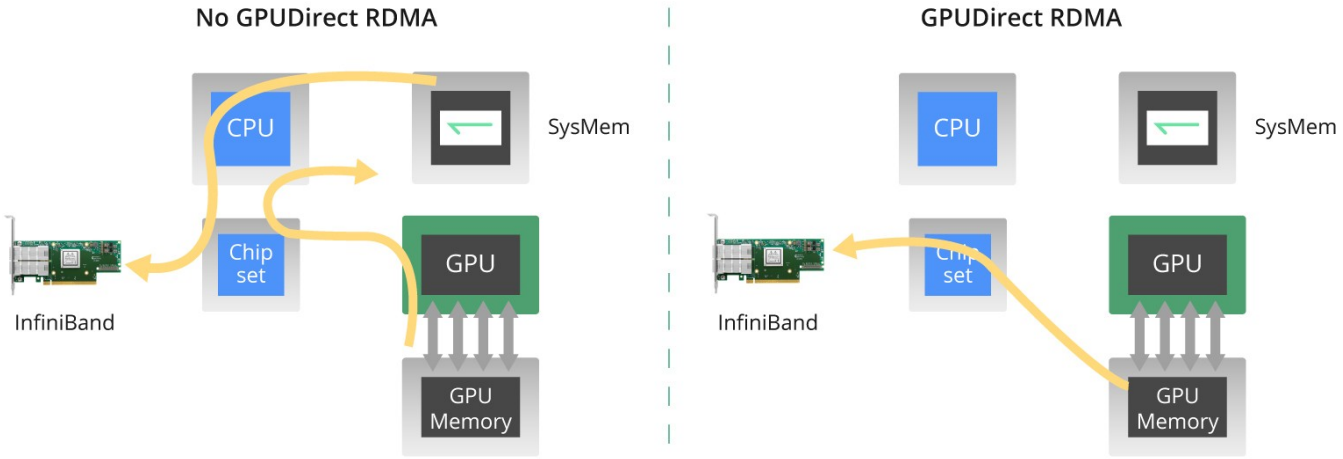
DESMOS SUPERCOMPUTER

Desmos supercomputer

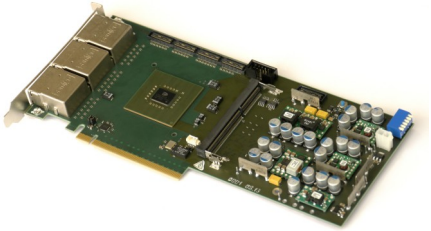


GPU-AWARE COMMUNICATIONS

NVIDIA GPUDirect RDMA technology for InfiniBand



A similar technology is under development for the Angara interconnect and is being tested with AMD GPUs



Preliminary results: **Khalilov M., Timofeev A., Polyakov D., Proceedings of RSD-2022, LNCS, vol. 13708, 2022.**

Micro-benchmarks, target application and analysis

OSU Micro-Benchmarks

ROCm/**rocHPL**

High Performance Linpack for Next-Generation AMD HPC Accelerators

AMD ROCm

Score-P

Scalable performance measurement infrastructure for parallel codes

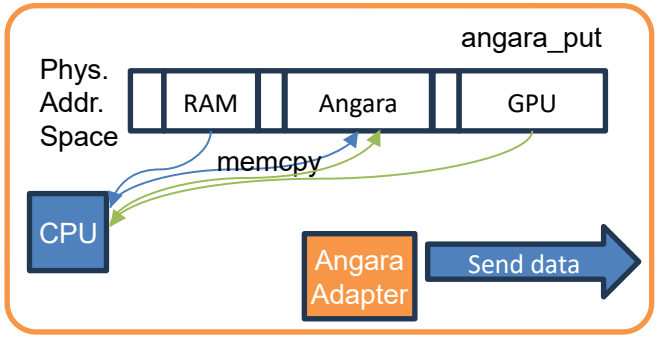
VAMPIR

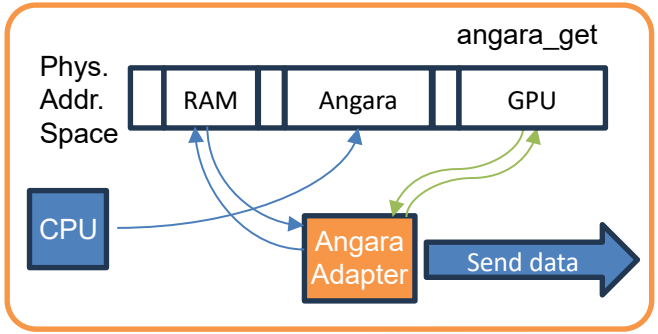
Angara base operations

- Each process using the Angara interconnect is numbered consecutively starting from zero and called the **pe**
- Angara interconnect support two base operations:
  - `angara_put(const u64_t *data, u64_t wroff, size_t len, u32_t dst)`
    - Copy data from **data* to the Angara injection buffer and send it to the *dst* by *wroff*, where wroff is the receive buffer offset in physical address space
  - `angara_get(u64_t wroff, u64_t rdoff, size_t len, u32_t src, u32_t dst)`
    - Send special package, which initiate data transmission from *rdoffset* of *src* to the *wroffset* of *dst*
    - angara_offload_get – dst is equal to the **pe** of the process that sent the request
    - angara_offload_put – src is equal to the **pe** of the process that sent the request

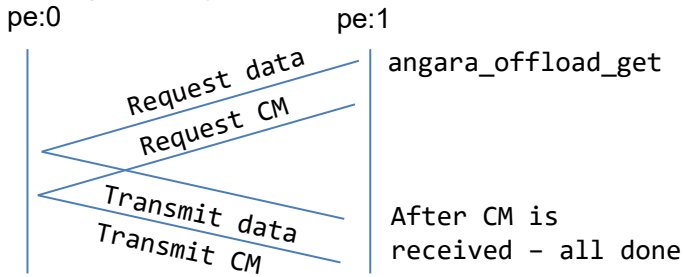
- For RDMA support is needing:
  - Data pointer translation to the physical memory address space mechanism
  - Data transfer confirmation mechanism

- Data transfer confirmation mechanism
  - Based on the property of the Angara GET packets that cannot "overtake" each other
  - In put_zcopy/get_zcopy scheme last GET packet send in the same way with predefined data (Completions Marker)

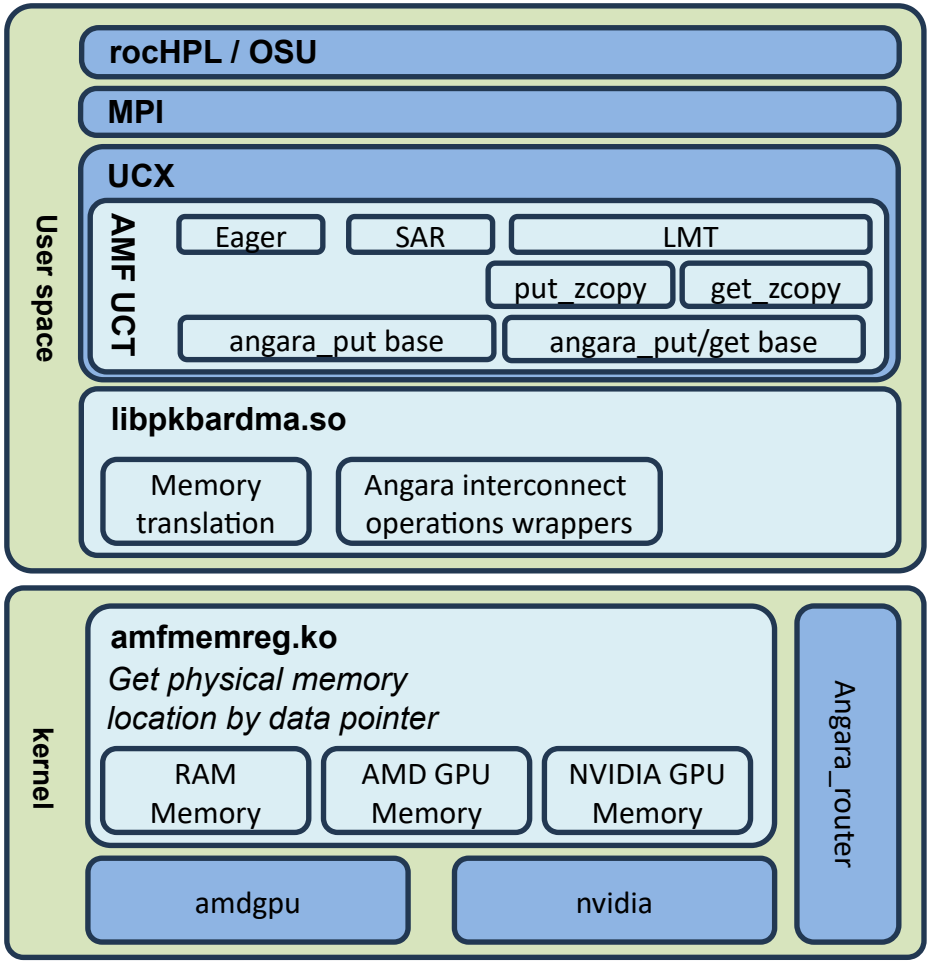




get_zcopy scheme



Angara API in UCX for GPU-aware MPI



AMF UCT – Transport in UCX implementing support for remote GPU or HOST memory reads and writes by extended Angara API.

libpkbardma.so – Extended Angara API for reads and writes operations and memory translation functions.

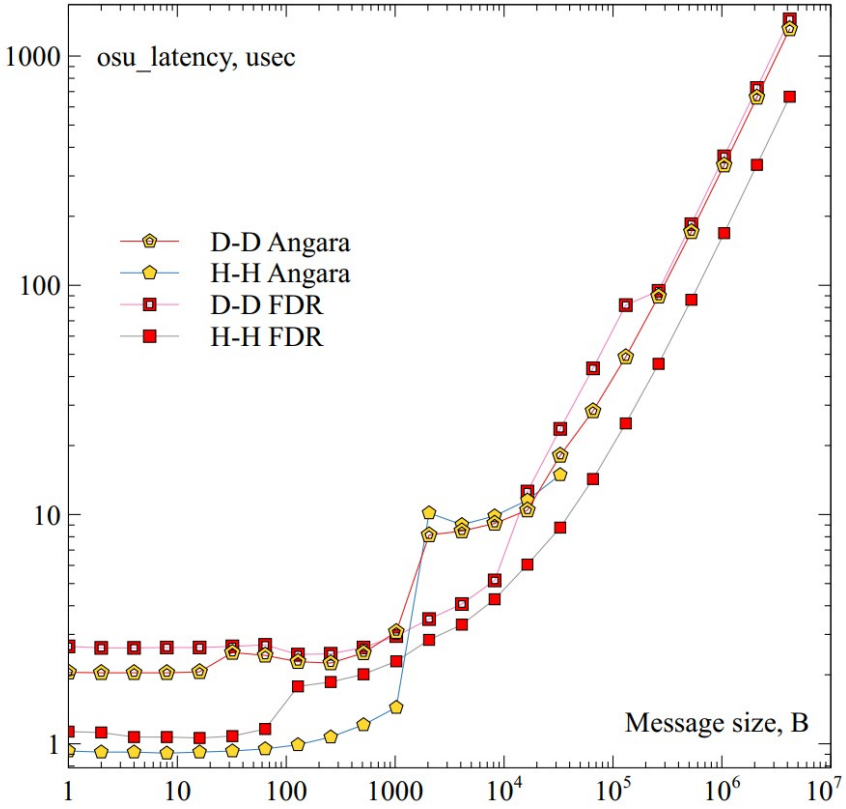
amfmemreg.ko – Kernel module for data pointer translation to the physical memory address space. Support:

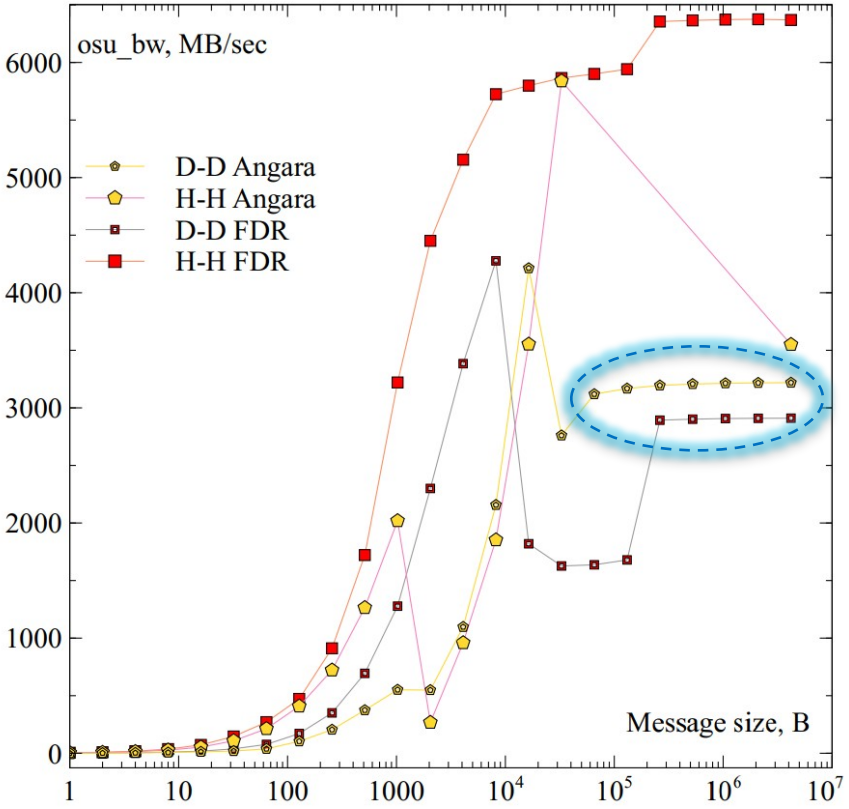
- HOST (RAM) Memory
- NVIDIA GPU Memory
- AMD GPU Memory

External modules/software

Developed modules/software

OSU micro-benchmarks: Angara vs InfiniBand FDR





rocHPL benchmark: Angara vs InfiniBand FDR with Score-P, obtained in May

Interconnect	MPI ranks	OMP threads	P	Q	N	with Score-P	Rmax GFlops
FDR	1	4	1	1	64512	no	4.2486e+03
FDR	1	4	1	1	64512	yes	4.2153e+03
Angara	1	4	1	1	64512	no	3.9523e+03
Angara	1	4	1	1	64512	yes	3.8281e+03
FDR	2	4	1	2	90624	no	8.0394e+03
FDR	2	4	1	2	90624	yes	7.9619e+03
Angara	2	4	1	2	90624	no	5.7822e+03
Angara	2	4	1	2	90624	yes	5.5953e+03
FDR	4	1	4	1	122000	no	1.4299e+04
FDR	4	1	4	1	122000	yes	1.4252e+04
Angara	4	1	4	1	122000	no	1.4571e+04
Angara	4	1	4	1	122000	yes	1.4511e+04

suboptimal variant due to host memory fragmentation problems

Rmax=53% of Rpeak

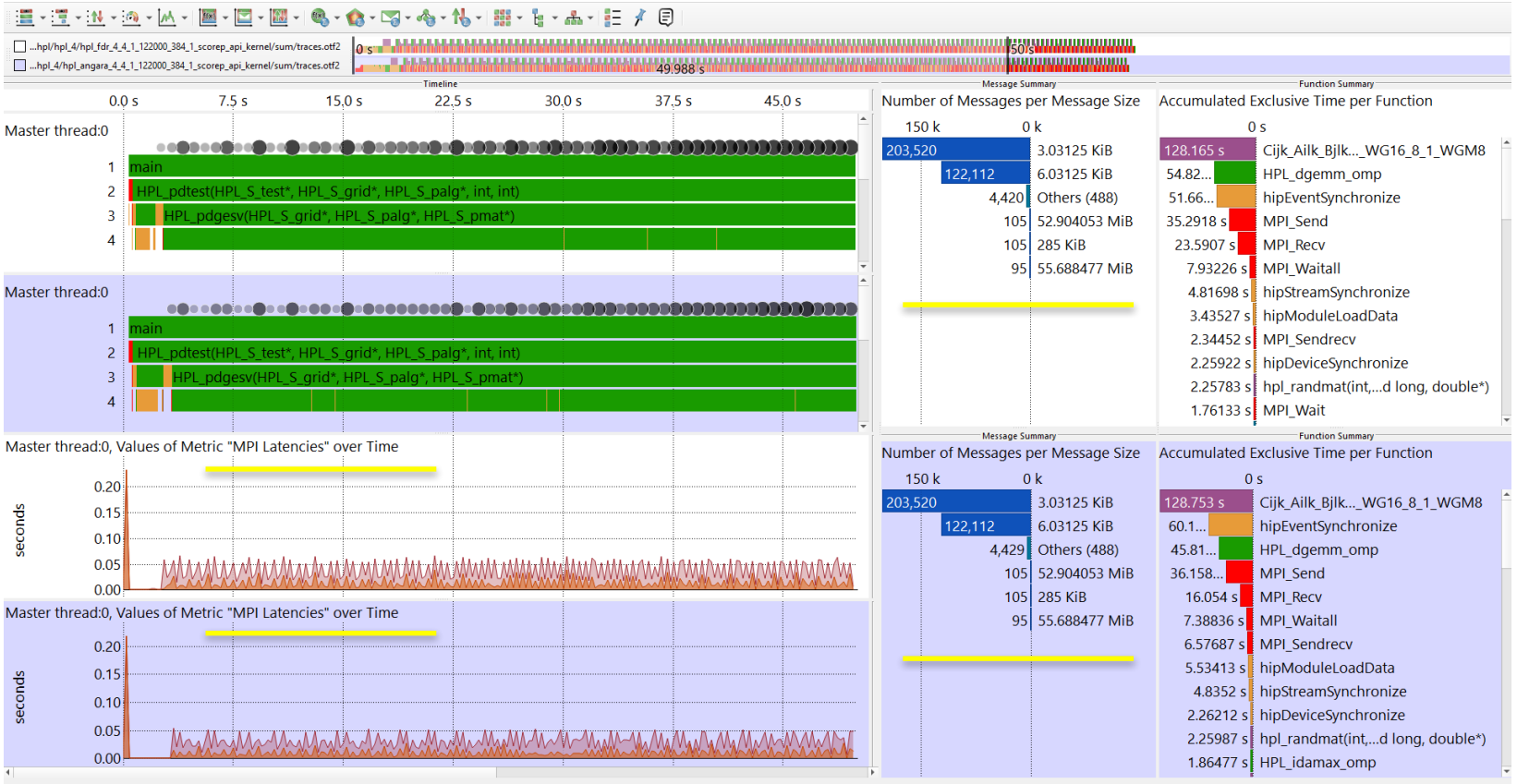
rocHPL benchmark: Angara
without Score-P, obtained in September

N	P	Q	VRAM	GFLOPS	
45312	1	1	51%	4.7003e+03	
54912	1	1	74%	4.7815e+03	
63360	1	1	97%	4.8252e+03	70% of Rpeak
N	**P**	**Q**	**VRAM**	**GFLOPS**	
45312	1	2	27%	7.7335e+03	
63360	1	2	51%	8.7673e+03	
89472	1	2	98%	9.1389e+03	
N	**P**	**Q**	**VRAM**	**GFLOPS**	
63360	2	2	27%	1.4431e+04	
89472	2	2	51%	1.6486e+04	
124800	2	2	96%	1.6891e+04	61% of Rpeak
N	**P**	**Q**	**VRAM**	**GFLOPS**	
89472	2	4	42%	2.2392e+04	
124800	2	4	65%	2.8304e+04	
158208	2	4	88%	3.2262e+04	
N	**P**	**Q**	**VRAM**	**GFLOPS**	
196992	4	4	61%	6.2011e+04	
230400	4	4	82%	6.2826e+04	57% of Rpeak

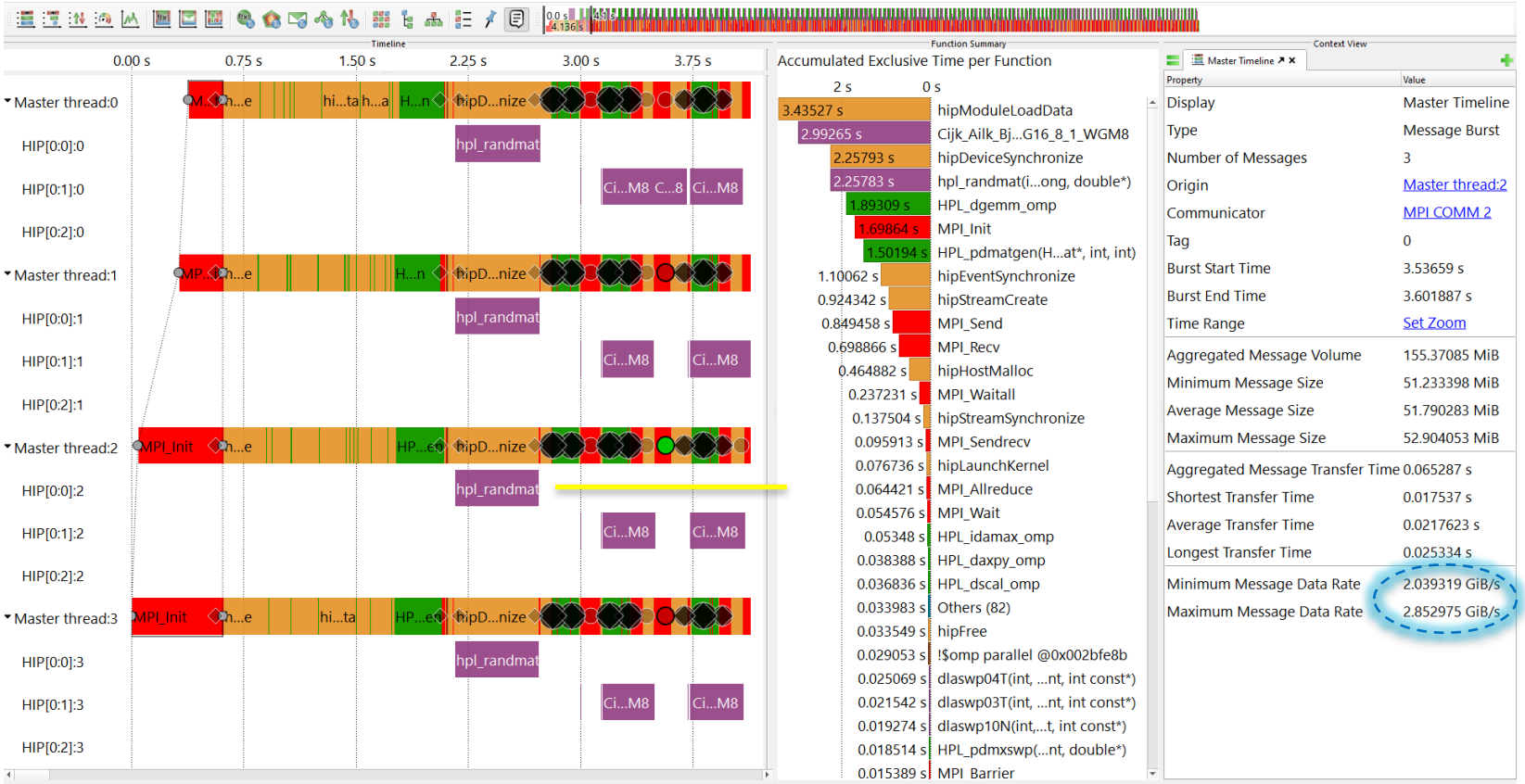
rocHPL benchmark: Angara
without Score-P, obtained in September

Interconnect	N	P	Q	VRAM	GFLOPS	
Angara	311040	6	5	80%	1.1271e+05	54% of Rpeak
FDR	311040	6	5	80%	1.1285e+05	
Angara	322560	6	5	88%	1.1361e+05	55% of Rpeak
FDR	322560	6	5	88%	1.1368e+05	

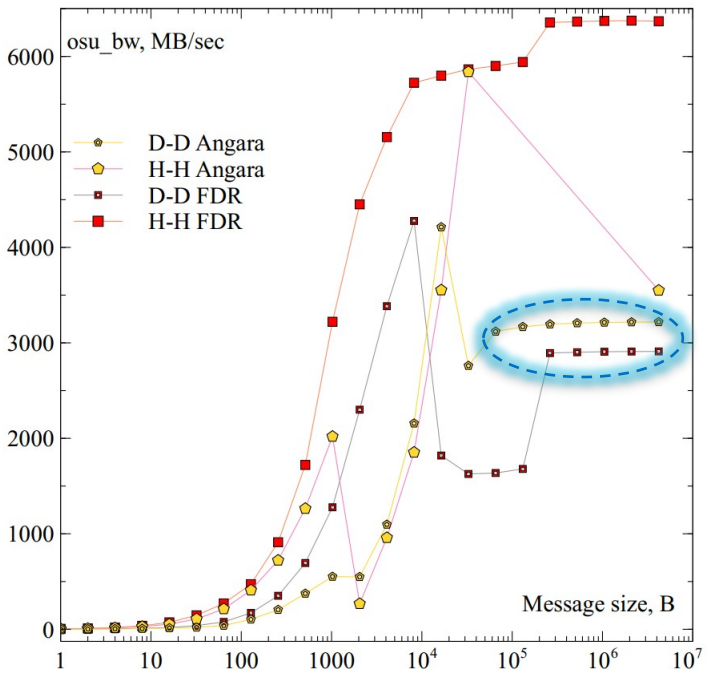
rocHPL benchmark: Angara vs InfiniBand FDR



rocHPL benchmark: InfiniBand FDR data rate



rocHPL benchmark: Angara data rate



Заключение

Russian Science Foundation
Grant No.20-71-10127

- Была представлена технология GPU-aware MPI для Ангары, основанная на программном модуле UCX RDMA.
- Были сравнены задержки и пропускные способности точка-точка коммуникаций Ангары с аналогичными у InfiniBand FDR.
- Для Ангары была измерена производительность rocHPL с различными комбинациями параметров и сравнена с производительностью InfiniBand FDR.
- Бенчмарк rocHPL был инструментирован с помощью Score-P, и его исполнение было проанализировано для Ангары и InfiniBand FDR с точки зрения трассировок.
- Анализ показал минимальные вычислительные расходы на использование Score-P, и было обнаружено соответствие пропускных способностей на osu_bw пропускным способностям, полученным со Score-P.