E·C·E

European Commission

Horizon 2020
European Union funding
for Research & Innovation

WP4: I/O & Data Flow

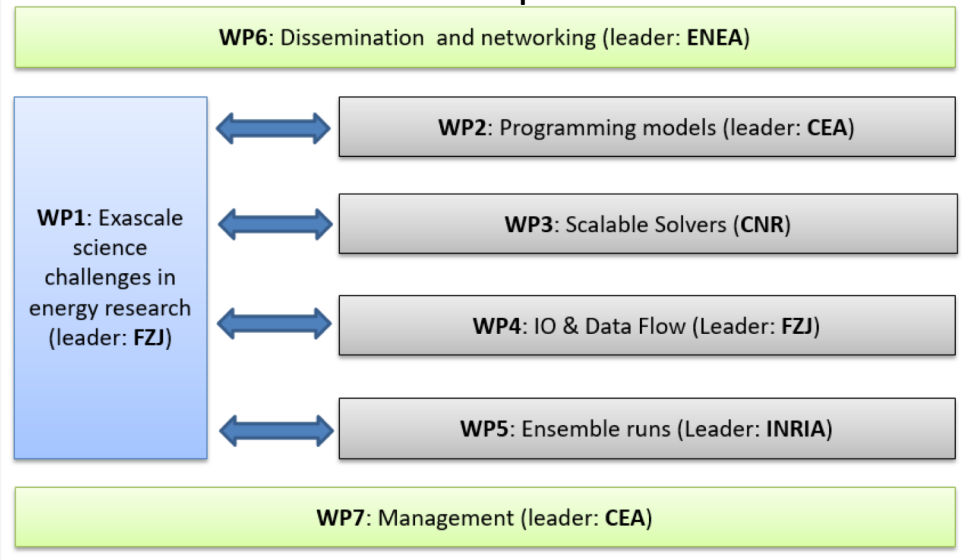
F.Iannone,F.Ambrosino,A.Funel, ENEA

Report Project – 12/02/2020

EoCoE^2: *Energy Oriented Center of Excellence : toward exascale for energy*

- ✓ HPC apps: Wind, Meteo, Materials, Water, Fusion
- ✓ Start date: 2019 – End Date: 2021

Work plan



PM efforts

	Partner number	WP1	WP2	WP3	WP4	WP5	WP6	WP7	Total Person-Months per Participant
CEA	1	56	70	2	24	30	5	30	217
FZJ	2	75	36	6	39	17	11	19	203
ENEA	3	30	0	0	7	0	25	1	63
BSC	4	23	42	12	27	0	1	1	106
CNRS	5	16	0	27	24	0	0	1	68
INRIA	6	14	14	18	2	40	1	1	90
CERFACS	7	12	0	33	0	0	1	1	47
MPG	8	39	15	2	0	0	1	1	58
FRAUNHOFER	9	24	0	0	0	0	1	1	26
FAU	10	0	30	0	0	0	1	1	32
CNR	11	0	0	48	0	0	0	0	48
UNITN	12	28	0	0	0	0	1	1	30
PSNC	13	0	0	0	48	18	24	0	90
ULB	14	0	0	33	0	0	1	1	35
UBAH	15	30	0	3	0	0	1	1	35
CIEMAT	16	7	0	0	0	0	2	1	10
IFPEN	17	3	18	0	0	0	1	1	23
DDN	18	0	0	0	24	0	1	1	26
TOTAL Person Months		357	225	184	195	105	78	63	1 207

- ✓ Project funds: 1207 PMs – 9.2 MEuro
- ✓ ENEA efforts : 63 PMs – 560 kEuro
- ✓ ENEA in T4.2: I/O Refactoring & optimization – T4.2.2: Gysela I/O optimization – 7 PMs

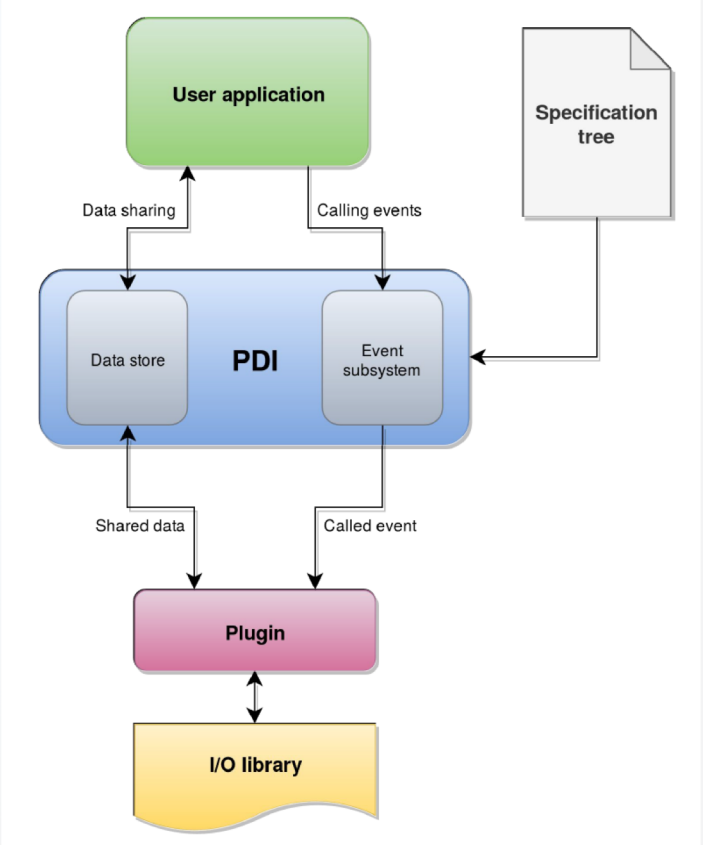
EoCoE^2: *Energy Oriented Center of Excellence : toward exascale for energy*

✓ WP 4 : I/O & Data Flow

- Task 4.2: I/O refactoring and optimization
  - Task 4.2.2: Gysela I/O optimization (Fusion) (Gysela-X not ready yet)

✓ PDI: Parallel Data Interface

- Middleware to interface I/O Library (Posix,HDF5,SionLib)
- C++/MPI comm
- Parser YAML allows to define data structure
- API:
  - PDI_init
  - PDI_share
  - PDI_reclaim
  - PDI_release
  - PDI_expose
  - PDI_access
  - PDI_event
  - PDI_multiexpose
  - PDI_finalize

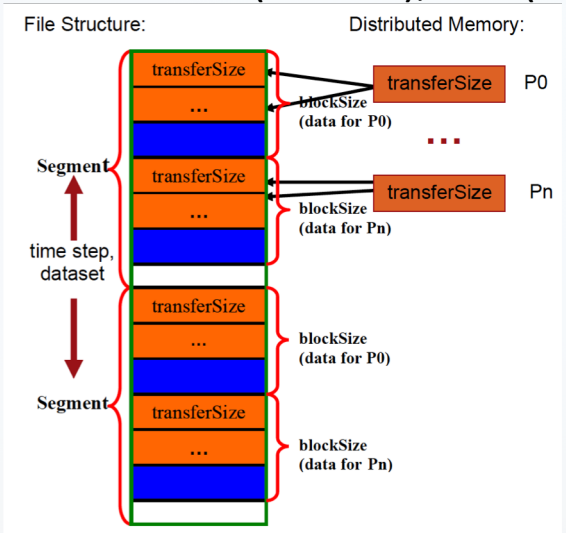


EoCoE^2: *Energy Oriented Center of Excellence : toward exascale for energy*

✓ **ENEA activities**:

- integrate PDI into the standard I/O benchmarking tool IOR waiting for GyselaX deployment
- test performances of PDI integrated in IOR with POSIX interface
- make an extensive benchmarks session using JUBE to automatically test, many parameters (e.g. file system block size, shared/independent access, collective I/O, cache effects etc.)
- find and report any bugs. Comparison of runs with and without PDI to give an estimation of the overhead especially for large I/O amount of data and/or large number of tasks

✓ Currently: installed standalone PDI (v. 0.5.1), IOR (v. 3.3.0) JUBE (v. 2.1.4).



IOR file structure and Processors

✓ API development in IOR of a PDI interface (POSIX plugin) in IOR: aiori-PDI.c

- PDI_create,PDI_Mknode,PDI_Open,**PDI_Xfer,** PDI_Close, PDI_Fsync, PDI_GetFileSize

EoCoE^2: *Energy Oriented Center of Excellence : toward exascale for energy*

✓ Plugin: xfer_pdi to interface PDI to POSIX functions read()/write()

```
int main(int argc,char *aa[])
 {
  int bs,ts,sc;
  int fd;
  int n_block;
  char *buffer;
  if(argc!=5){
       printf("Usage : filename blocksize[MB] transfersize[MB] segmentcount \n");
       return -1;
   }
  fd=open(aa[1],O_CREAT|O_RDWR);
   if(fd==-1){
       printf("file create error.\n");
       return -1;
   }
  bs=atoi(aa[2]);
  ts=atoi(aa[3]);
  sc=atoi(aa[4]);
  printf("bs: %d\n",bs);
  n_block = bs/ts;
  buffer = malloc(sizeof(char)*ts*1024*1024);
  memset(buffer,1,ts*1024*1024);
  for (int i = 0; i < sc; i++) {
     for (int j = 0; j < n_block; j++) {
        printf("block: %d %d\n",i,j);
        write(fd,buffer,ts*1024*1024);
     }
  }

close(fd);

}
```

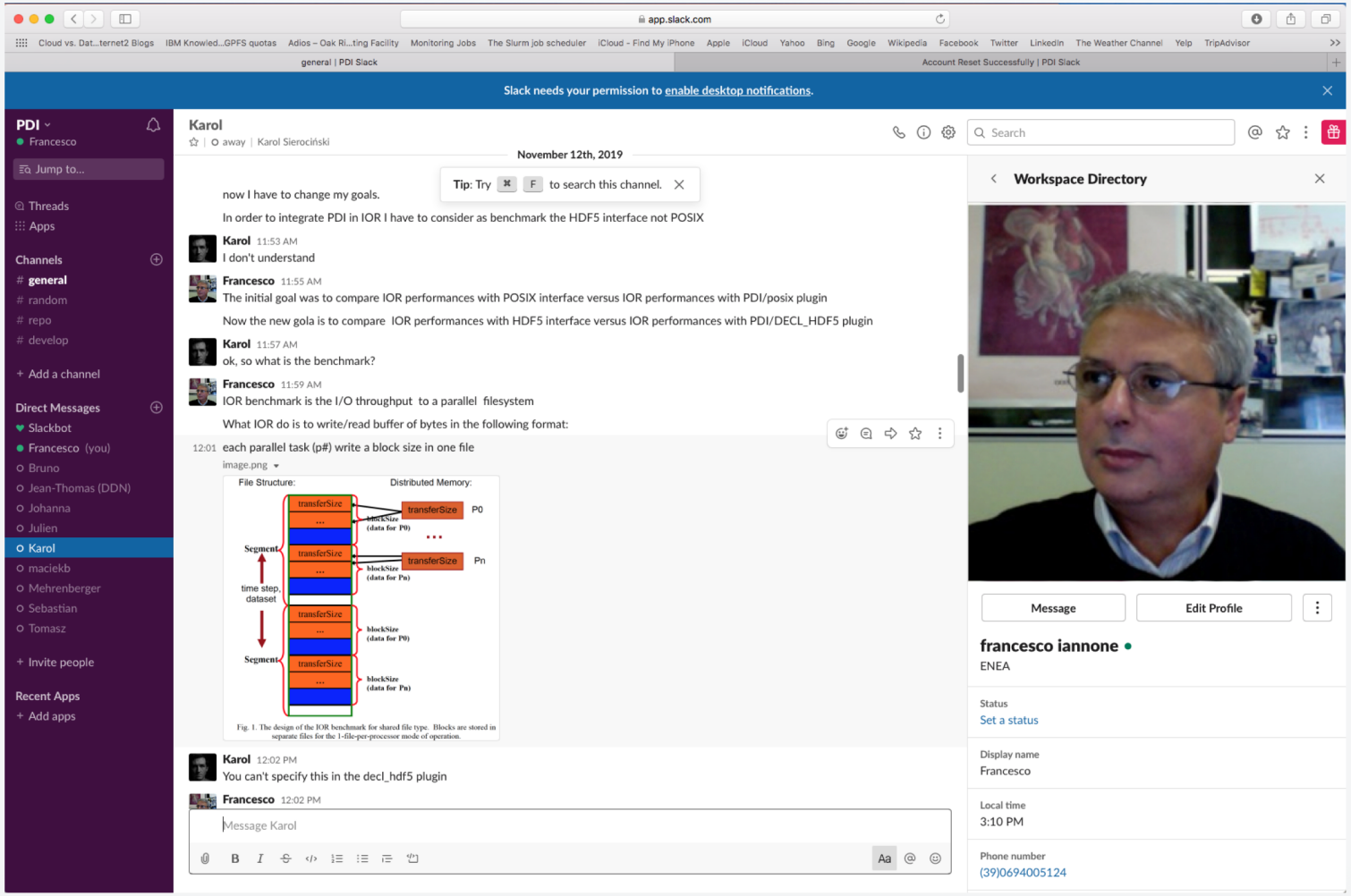
IOR POSIX benchmark

```
YAML file
metadata:
  ts: int64
  access: int
  file: int
  offset: int64
  return: int64
data:
  buffer: {type: array, subtype: char, size: $ts}
plugins:
  user_code:
    on_data:
      buffer:
        xfer_pdi:
          value: $buffer
```

```
133 void xfer_pdi()
134 {
135     // arguments of the function
136     int* access;  PDI_access("access", (void**)&access, PDI_IN);
137     long* length; PDI_access("ts", (void**)&length, PDI_IN);
138     long* offset; PDI_access("offset", (void**)&offset, PDI_IN);
139     int* fd;      PDI_access("file", (void**)&fd, PDI_IN);
140     long l=*length;
141     char* pptr;
142     if (*access == READ) { //read
143         PDI_access("value", (void**)&pptr, PDI_OUT);
144     } else { // write
145         PDI_access("value", (void**)&pptr, PDI_IN);
146     }
147     long long rc;
148     long long remaining = (long long)*length;
149     // function body
150     if (*access == READ) {
151         if (verbose >= VERBOSE_4) {
152                                 fprintf(stdout,
153                                         "task %d reading from offset %lld\n",
154                                         rank,
155                                         *offset + length - remaining);
156                     }
157         rc = read(*fd, pptr, *length);
158     } else {
159         if (verbose >= VERBOSE_4) {
160                                 fprintf(stdout,
161                                         "task %d writing to offset %lld\n",
162                                         rank,
163                                         *offset + length - remaining);
164                     }
165         rc = write(*fd, pptr, *length);
166
167     }
168
169     // release all metadata
170     PDI_release("access");
171     PDI_release("file");
172     PDI_release("ts");
173     PDI_release("offset");
174     PDI_release("value");
175     // update return value
176     PDI_expose("return", &rc, PDI_OUT);
177
178 }
```

EoCoE^2: *Energy Oriented Center of Excellence : toward exascale for energy*

✓ Collaborative work: SLACK – PDI *https://app.slack.com/client/T9G3CB93N/DL5P6H8KY/user_profile/UL3537YHJ*



EoCoE^2: *Energy Oriented Center of Excellence : toward exascale for energy*

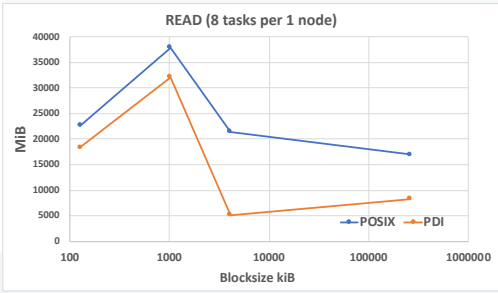
✓ TEST ENVIRONMENT: Jube

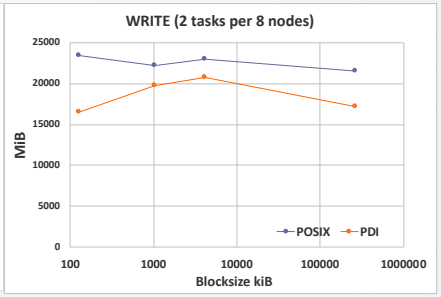
IORPDI.xml

```
<parameter name="api">POSIX,PDI</parameter>
    <parameter name="blockSize">${transferSize}</parameter>
    <parameter name="transferSize" type="int" mode="python">",".join(str(i) for i in [128*1024,1024**2,4*(1024**2),256*(1024**2)])</parameter>
    <parameter name="segmentCount" type="int" mode="python">(2*(256*1024**2))/$transferSize</parameter>
    <parameter name="repetitions" type="int">3</parameter>
        <parameter name="filePerProc" type="int">1</parameter>
          <parameter name="keepFile" type="int">1</parameter>
          <parameter name="testFile">${workdir}/testFile</parameter>
        <parameter name="workdir">/gporq3/eocoe/work/$jube_wp_id</parameter>  <!--originariamente era indicata come testdir-->
        <parameter name="nodes" type="int">1,2,4,6,8</parameter>
        <parameter name="taskspernode" type="int">1,2,4,8,16,32</parameter>
        <parameter name="tasks" type="int" mode="python">$nodes*$taskspernode</parameter>
```

```
(start) 2019-12-06 16:29:35 (1575646175)
IOR-3.3.0+dev: MPI Coordinated Test of Parallel I/O
Began               : Fri Dec  6 16:29:35 2019
Command line        : /gporq2/store_1/project/eocoe2/iorpdi/ior/build_openmpi_gcc730/bin/ior -f ior_input.cfg -o /gporq3/eocoe/work/24 -k
Machine             : Linux cresco6x293.portici.enea.it 3.10.0-514.26.2.el7.x86_64 #1 SMP Tue Jul 4 15:04:05 UTC 2017 x86_64
Using synchronized MPI timer
Start time skew across all tasks: 0.00 sec
TestID              : 0
StartTime           : Fri Dec  6 16:29:35 2019
Path                : /gporq3/eocoe/work/24
FS                  : 756.0 TiB   Used FS: 33.9%   Inodes: 1918.8 Mi   Used Inodes: 0.3%
Participating tasks: 1
Using reorderTasks '-C' (expecting block, not cyclic, task assignment)

Options:
api                 : PDI
apiVersion          :
test filename       : /gporq3/eocoe/work/24/testFile
access              : file-per-process
type                : independent
segments            : 4096
ordering in a file  : sequential
ordering inter file : constant task offset
task offset         : 1
tasks               : 1
clients per node    : 1
repetitions         : 10
xfersize            : 131072 bytes
blocksize           : 131072 bytes
aggregate filesize  : 512 MiB

Results:
Using Time Stamp 1575646175 (0x5dea73df) for Data Signature

access    bw(MiB/s)  block(KiB) xfer(KiB)  open(s)    wr/rd(s)   close(s)  total(s)  iter
------    ---------  ---------- ---------  --------   --------   --------  --------  ----
[PDI][User-code][16:29:35] *** info: Plugin loaded successfully
[PDI][16:29:35] *** info: Initialization successful
Commencing write performance test: Fri Dec  6 16:29:35 2019
[PDI][16:29:36] *** info: Finalization
[PDI][User-code][16:29:36] *** info: Closing plugin
write     1478.95    128.00     128.00     0.008377   0.337223   0.000546   0.346191   0
read      2711.39    2670.33    2689.65    14.15      21691.11   21362.63   21517.21   113.23   0.19036    0   1   1   10   1   1   1   0   0   4096   131072   131072   512.0 PDI   9999

Operation  Max(MiB)  Min(MiB)  Mean(MiB)   StdDev  Max(OPs)  Min(OPs)  Mean(OPs)   StdDev   Mean(s) Test# #Tasks tPN reps fPP reord reordoff reordrand seed segcnt  blksiz  xsize aggs(MiB)  API RefNum
write       1963.58   1478.95   1849.75   128.16  15708.64  11831.60  14798.00  1025.24  0.27839    0   1   1   10   1   1   1   0   0   4096   131072   131072   512.0 PDI   9999
read        2711.39   2670.33   2689.65    14.15  21691.11  21362.63  21517.21   113.23  0.19036    0   1   1   10   1   1   1   0   0   4096   131072   131072   512.0 PDI   9999

Summary of all tests:
Operation  Max(MiB)  Min(MiB)  Mean(MiB)   StdDev  Max(OPs)  Min(OPs)  Mean(OPs)   StdDev   Mean(s) Test# #Tasks tPN reps fPP reord reordoff reordrand seed segcnt  blksiz  xsize aggs(MiB)  API RefNum
write       1963.58   1478.95   1849.75   128.16  15708.64  11831.60  14798.00  1025.24  0.27839    0   1   1   10   1   1   1   0   0   4096   131072   131072   512.0 PDI   9999
read        2711.39   2670.33   2689.65    14.15  21691.11  21362.63  21517.21   113.23  0.19036    0   1   1   10   1   1   1   0   0   4096   131072   131072   512.0 PDI   9999
Finished            : Fri Dec  6 16:29:40 2019
```

nodes	taskspernode	tasks	API	RCBlock[KiB]	RDtw[MiB]	RDclose[s]	RDopen[s]	RDrd[s]	RDtotal[s]	RDxfer[KiB]	WRblock[KiB]	WRtw[MiB]	WRclose[s]	WRopen[s]	WRtotal[s]	WRwr[s]	WRxfer[KiB]
1	1	1	POSIX	128	4916	7.00E-05	2.30E-05	0.104046	0.104155	128	128	2500.2	0.000192	0.002295	0.104783	0.002267	128
1	1	1	POSIX	1024	4976	9.00E-06	2.60E-05	0.102837	0.102889	1024	1024	1861.79	0.000225	0.02336	0.275004	0.251374	1024
1	1	1	POSIX	4096	4916	5.10E-05	2.70E-05	0.103214	0.103309	4096	4096	2137.99	0.000237	0.000867	0.239477	0.238133	4096
1	1	1	POSIX	262144	3667	0.000119	3.40E-05	0.139442	0.139611	262144	262144	2066.67	0.000163	0.000107	0.247742	0.247333	262144
1	2	2	POSIX	128	11555	1.30E-05	3.70E-05	0.088559	0.088623	128	128	4152	0.000208	0.002408	0.246648	0.240912	128
1	2	2	POSIX	1024	10589	2.50E-05	3.90E-05	0.096624	0.096704	1024	1024	3820	0.000367	0.002466	0.26804	0.265193	1024
1	2	2	POSIX	4096	10516	0.000139	3.40E-05	0.096822	0.097009	4096	4096	4031	0.000298	0.002009	0.25378	0.253147	4096
1	2	2	POSIX	262144	7699	0.000203	3.00E-05	0.132766	0.133007	262144	262144	3212.38	0.000284	0.002349	0.318767	0.316122	262144
1	4	4	POSIX	128	19058	2.50E-05	4.10E-05	0.107384	0.10746	128	128	5776	0.000462	0.005313	0.35456	0.348757	128
1	4	4	POSIX	1024	21941	2.00E-05	3.80E-05	0.093267	0.093343	1024	1024	6754	0.000465	0.003821	0.303209	0.301801	1024
1	4	4	POSIX	4096	14138	0.00025	5.90E-05	0.144535	0.144858	4096	4096	5426	0.000429	0.002806	0.377468	0.374212	4096
1	4	4	POSIX	262144	14375	0.000473	6.80E-05	0.141912	0.142473	262144	262144	6498	0.000438	0.003629	0.315174	0.311047	262144
1	8	8	POSIX	128	25223	3.40E-05	7.80E-05	0.162261	0.162394	128	128	5520	0.001002	0.013986	0.742053	0.727021	128
1	8	8	POSIX	1024	36998	4.10E-05	6.90E-05	0.110602	0.11071	1024	1024	5771	0.000918	0.00968	0.706481	0.698867	1024
1	8	8	POSIX	4096	13988	0.00071	7.80E-05	0.291101	0.29283	4096	4096	2977.93	0.001	0.007799	1.38	1.37	4096
1	8	8	POSIX	262144	14497	0.000926	7.40E-05	0.281558	0.282536	262144	262144	6827	0.000977	0.006035	0.599964	0.592971	262144
1	16	16	POSIX	128	3961	0.000187	0.00022	2.07	2.07	128	128	4733	0.001996	0.010204	1.73	1.72	128
1	16	16	POSIX	1024	4370	0.000544	0.000145	1.87	1.87	1024	1024	5737	0.000599	0.011775	1.43	1.42	1024
1	16	16	POSIX	4096	4122	0.000725	0.000117	1.99	1.99	4096	4096	6058	0.003403	0.010049	1.35	1.34	4096
1	16	16	POSIX	262144	4031	0.001844	0.000108	2.03	2.03	262144	262144	6325	0.00035	0.011299	1.3	1.28	262144
1	32	32	POSIX	128	1240.7	0.001448	0.000365	13.2	13.21	128	128	4784	0.001766	0.020629	3.42	3.4	128
1	32	32	POSIX	1024	1470.61	0.001627	0.000519	11.14	11.14	1024	1024	5261	0.001845	0.02049	3.11	3.09	1024
1	32	32	POSIX	4096	1546.12	0.002449	0.000393	10.59	10.6	4096	4096	5331	0.079432	0.02129	3.07	2.97	4096
1	32	32	POSIX	262144	1456.06	0.003898	0.0005	11.25	11.25	262144	262144	4283	0.003552	0.036057	3.83	3.79	262144
1	1	1	PDI	128	2887.73	0.000213	0.00035	0.176725	0.177302	128	128	1665.36	0.000532	0.035218	0.307441	0.271667	128
1	1	1	PDI	1024	5494	0.000242	0.000359	0.09255	0.093199	1024	1024	2812.29	0.00056	0.003023	0.182058	0.178453	1024
1	1	1	PDI	4096	6256	0.000274	0.000372	0.079895	0.080555	4096	4096	3109.84	0.000472	0.002313	0.164639	0.16183	4096
1	1	1	PDI	262144	4451	0.000354	0.000439	0.114235	0.115061	262144	262144	2217.32	0.00046	0.008035	0.23091	0.222388	262144
1	2	2	PDI	128	5426	0.000324	0.000397	0.187974	0.188711	128	128	2813.7	0.000618	0.010352	0.364182	0.3532	128
1	2	2	PDI	1024	10041	0.000311	0.000435	0.101217	0.101978	1024	1024	4492	0.000552	0.003065	0.227985	0.224315	1024
1	2	2	PDI	4096	5383	0.000468	0.000377	0.189377	0.190229	4096	4096	3478.18	0.000551	0.024577	0.294407	0.269254	4096
1	2	2	PDI	262144	8275	0.00045	0.000393	0.122882	0.123739	262144	262144	4572	0.000614	0.008111	0.223985	0.215228	262144
1	4	4	PDI	128	9851	0.000346	0.000541	0.206995	0.207893	128	128	4990	0.00074	0.003929	0.410384	0.405696	128
1	4	4	PDI	1024	18145	0.000292	0.000539	0.112032	0.112871	1024	1024	6242	0.000823	0.010157	0.328117	0.317097	1024
1	4	4	PDI	4096	4165	0.000473	0.000529	0.490652	0.491666	4096	4096	5459	0.000689	0.039956	0.37515	0.33448	4096
1	4	4	PDI	262144	14707	0.000775	0.000759	0.137712	0.139257	262144	262144	6090	0.000704	0.030483	0.336308	0.305044	262144
1	8	8	PDI	128	20273	0.001181	0.001356	0.199523	0.202041	128	128	5823	0.001191	0.010187	0.703366	0.691966	128
1	8	8	PDI	1024	35195	0.001063	0.001265	0.11405	0.116379	1024	1024	5609	0.001249	0.031856	0.780366	0.697136	1024
1	8	8	PDI	4096	5399	0.001234	0.000837	0.75057	0.758059	4096	4096	5788	0.001509	0.018819	0.707673	0.687323	4096
1	8	8	PDI	262144	12789	0.000858	0.002365	0.31609	0.320279	262144	262144	4634	0.001979	0.027313	0.883817	0.854492	262144
1	16	16	PDI	128	4246	0.001844	0.00257	1.93	1.93	128	128	4458	0.002567	0.007968	1.84	1.77	128
1	16	16	PDI	1024	4644	0.002455	0.002544	1.76	1.76	1024	1024	5675	0.002584	0.01549	1.44	1.43	1024
1	16	16	PDI	4096	4043	0.002258	0.002416	2.02	2.03	4096	4096	5989	0.002365	0.048549	1.37	1.32	4096
1	16	16	PDI	262144	3900	0.003438	0.00241	2.05	2.05	262144	262144	6234	0.002392	0.043948	1.31	1.27	262144
1	32	32	PDI	128	1295.43	0.002968	0.004785	12.64	12.65	128	128	4684	0.004339	0.034328	3.5	3.4	128
1	32	32	PDI	1024	1638.64	0.005669	0.005671	9.98	10	1024	1024	5203	0.005364	0.062078	3.15	3.08	1024
1	32	32	PDI	4096	1377.97	0.006677	0.005896	11.88	11.89	4096	4096	4994	0.015795	0.077296	3.28	3.19	4096
1	32	32	PDI	262144	1359.19	0.004226	0.005565	12.04	12.05	262144	262144	4990	0.009919	0.075127	3.28	3.2	262144





- ✓ **Next actions**:
  - New IOR POSIX/PDI benchmarks in a noiseless sessions
  - Deliverable of WP4 with ENEA contribute on PDI integration
  - Development of Gysela optimization (coordinate by JSC)
  - Effort need: 1 developer for Gysela optimization