Outro To Parallel Computing

John Urbanic
Pittsburgh Supercomputing Center
Parallel Computing Scientist
Purpose of this talk

Now that you know how to do some real parallel programming, you may wonder how much you don’t know. With your newly informed perspective we will take a look at the parallel software landscape so that you can see how much of it you are equipped to traverse.
How parallel is a code?

- Parallel performance is defined in terms of scalability

**Strong Scalability**
Can we get faster for a given problem size?

**Weak Scalability**
Can we maintain runtime as we scale up the problem?
Weak vs. Strong scaling

Weak Scaling:
- More Processors
- More accurate results

Strong Scaling:
- More Processors
- Faster results

(Tornado on way!)
How many processors can we really use?

Let’s say we have a legacy code such that it is only feasible to convert half of the heavily used routines to parallel:
Amdahl’s Law

If we run this on a parallel machine with five processors:

Our code now takes about 60s. We have sped it up by about 40%.

Let’s say we use a thousand processors:

We have now sped our code by about a factor of two. Is this a big enough win?
Amdahl's Law

- If there is x% of serial component, speedup cannot be better than $100/x$.

- If you decompose a problem into many parts, then the parallel time cannot be less than the largest of the parts.

- If the critical path through a computation is $T$, you cannot complete in less time than $T$, no matter how many processors you use.

- Amdahl's law used to be cited by the knowledgeable as a limitation.

- These days it is mostly raised by the uninformed.

- Massive scaling is commonplace:
  - Science Literature
  - Web (map reduce everywhere)
  - Data Centers (Spark, etc.)
  - Machine Learning (GPUs and others)
Need to write some scalable code?

First Choice:

Pick a language - or maybe a library, or paradigm (whatever that is)?
Languages: Pick One (Hint: MPI + ?)

Parallel Programming environments since the 90’s

<table>
<thead>
<tr>
<th>ABCPL</th>
<th>CORRELATE</th>
<th>GLU</th>
<th>Mentat</th>
<th>Parafine2</th>
<th>pC++</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACI</td>
<td>CPS</td>
<td>GUARD</td>
<td>Legion</td>
<td>Parallel</td>
<td>SCHEDULE</td>
</tr>
<tr>
<td>ACT++</td>
<td>CRL</td>
<td>HAsL</td>
<td>Meta Chaos</td>
<td>Parallel-C++</td>
<td>ScTIL</td>
</tr>
<tr>
<td>Active messages</td>
<td>CSP</td>
<td>Haskell</td>
<td>Midway</td>
<td>Parallaxis</td>
<td>POET</td>
</tr>
<tr>
<td>Adl</td>
<td>Chthreads</td>
<td>HPC++</td>
<td>Multipede</td>
<td>ParC</td>
<td>SDDA</td>
</tr>
<tr>
<td>Adsmith</td>
<td>CUMULVS</td>
<td>JAVAR</td>
<td>CpaPar</td>
<td>ParLib++</td>
<td>SIMD</td>
</tr>
<tr>
<td>ADDAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFAPI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALWAN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMDC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AppleC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARTS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atlastic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aurora</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automap</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bthreads</td>
<td>Blaze</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BSP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BlockComm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C++</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C* in C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C却又</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cashmere</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC++</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charlotte</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chro</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chro++</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ctd</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clik</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CM-Fortran</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Converse</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Code</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COOL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAMMA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glenda</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lpars</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Local</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Massic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manifold</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pages</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAPI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Para++</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paradigm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quick</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Threads</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sage++</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCANDAL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Alternative Approach

“When all you have is a hammer, everything looks like a nail.”
Paradigm?

- Message Passing
  - MPI
- Data Parallel
  - Fortran90
- Threads
  - OpenMP, OpenACC, CUDA
- PGAS
  - UPC, Coarray Fortran
- Frameworks
  - Charm++
- Hybrid
  - MPI + OpenMP
Message Passing: MPI in particular

Pros

- Has been around a longtime (~20 years inc. PVM)
- Dominant
- Will be around a longtime (on all new platforms/roadmaps)
- Lots of libraries
- Lots of algorithms
- Very scalable (100K+ cores right now)
- Portable
- Works with hybrid models
- We teach MPI in two days also
- This is the only route to massive scalability today!

Cons

- Lower level means more detail for the coder
- Debugging requires more attention to detail
- Domain decomposition and memory management must be explicit
- Students leaving our MPI workshop may face months of work before they are able to actually run their production code
- Development usually requires a “start from scratch” approach
Data Parallel – Fortran90

Computation in FORTRAN 90

Real A(4,4), B(4,4), C(4,4)

A=2.0
FORALL (I=1:4, J=1:4)
  B(I,J)=I+J
C=A+B

P = Processor

\[
\begin{array}{cccc}
4 & 5 & 6 & 7 \\
5 & 6 & 7 & 8 \\
6 & 7 & 8 & 9 \\
7 & 8 & 9 & 10 \\
\end{array}
\]

\[
\begin{array}{cccc}
2 & 2 & 2 & 2 \\
2 & 2 & 2 & 2 \\
2 & 2 & 2 & 2 \\
2 & 2 & 2 & 2 \\
\end{array}
\]

\[
\begin{array}{cccc}
2 & 3 & 4 & 5 \\
3 & 4 & 5 & 6 \\
4 & 5 & 6 & 7 \\
5 & 6 & 7 & 8 \\
\end{array}
\]
Data Parallel Communication in FORTRAN 90

<table>
<thead>
<tr>
<th>P</th>
<th>P</th>
<th>P</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
</tbody>
</table>

Real $A(4,4), B(4,4)$

FORALL (I=1:4, J=1:4)

$B(I, J)=I+J$

$A=CSHIFT (B, DIM=2, 1)$

$P = Processor$

\[\begin{array}{cccc}
3 & 4 & 5 & 6 \\
4 & 5 & 6 & 7 \\
5 & 6 & 7 & 8 \\
2 & 3 & 4 & 5 \\
\end{array}\]  \hspace{2cm} \begin{array}{cccc}
2 & 3 & 4 & 5 \\
3 & 4 & 5 & 6 \\
4 & 5 & 6 & 7 \\
5 & 6 & 7 & 8 \\
\end{array}\]
Data Parallel

Pros

- So simple you just learned some of it
- ...or already knew it from using Fortran
- Easy to debug

Cons

- If your code doesn’t totally, completely express itself as nice array operations, you are left without a flexible alternative.
  - Image processing: Great
  - Irregular meshes: Not so great
Threads in OpenMP

Fortran:

```fortran
!$omp parallel do
do i = 1, n
    a(i) = b(i) + c(i)
enddo
```

C/C++:

```c
#pragma omp parallel for
for(i=1; i<=n; i++)
    a[i] = b[i] + c[i];
```
subroutine saxpy(n, a, x, y)
    real :: x(:), y(:), a
    integer :: n,
    i
!$acc kernels
do
i =1,n
y(i) = a*x(i)+y(i)
enddo
!$acc end kernels
end subroutine

... From main program
$ call SAXPY on 1M elements
call saxpy(1<<20, 2.0, x, y);
...

void saxpy(int n,
    float a,
    float *x,
    float *restrict y)
{
    #pragma acc kernels
    for (int i = 0; i < n; ++i)
y[i] = a*x[i] + y[i];
}

... Somewhere in main
// call SAXPY on 1M elements
saxpy(1<<20, 2.0, x, y);
...

SAXPY in C

SAXPY in Fortran

subroutine saxpy(n, a, x, y)
    real :: x(:), y(:), a
    integer :: n, i
!$acc kernels
do i=1,n
    y(i) = a*x(i)+y(i)
endo
!$acc end kernels
end subroutine saxpy

... From main program
$ call SAXPY on 1M elements
call saxpy(2**20, 2.0, x_d, y_d)
...

Threads in OpenACC
Threads without directives: CUDA

// Host code
int main(int argc, char** argv)
{
    // Allocate input vectors in host memory
    h_A = (float*)malloc(size);
    if (h_A == 0) Cleanup();
    h_B = (float*)malloc(size);
    if (h_B == 0) Cleanup();
    h_C = (float*)malloc(size);
    if (h_C == 0) Cleanup();

    // Initialize input vectors
    Init(h_A, N);
    Init(h_B, N);

    // Allocate vectors in device memory
    cudaMalloc((void**)&d_A, size);
    cudaMalloc((void**)&d_B, size);
    cudaMalloc((void**)&d_C, size);

    // Copy vectors to device memory
    cudaMemcpy(d_A, h_A, size, cudaMemcpyHostToDevice);
    cudaMemcpy(d_B, h_B, size, cudaMemcpyHostToDevice);

    // Run kernel
    int threadsPerBlock = 256;
    int blocksPerGrid = (N + threadsPerBlock - 1) / threadsPerBlock;
    VecAdd<<<blocksPerGrid, threadsPerBlock>>>(d_A, d_B, d_C, N);

    // Copy results from device memory to host memory
    cudaMemcpy(h_C, d_C, size, cudaMemcpyDeviceToHost);

    // GPU Code
    __global__ void VecAdd(const float* A, const float* B, float* C, int N)
    {
        int i = blockDim.x * blockIdx.x + threadIdx.x;
        if (i < N)
            C[i] = A[i] + B[i];
    }
Threads

Splits up tasks (as opposed to arrays in data parallel) such as loops amongst separate processors.

Do communication as a side effect of data loop distribution. Not an big issue on shared memory machines. Impossible on distributed memory.

Common Implementations:
- pthreads (original Unix standard)
- OpenMP
- OpenACC
- OpenCL (Khronos Group)
- DirectCompute (Microsoft)

Very C++ oriented:
- C++ AMP (MS/AMD)
- TBB (Intel C++ template library)
- Cilk (Intel, now in a gcc branch)

Pros:
1. Doesn’t perturb data structures, so can be incrementally added to existing serial codes.
2. Becoming fairly standard for compilers.

Cons:
1. Serial code left behind will be hit by Amdahl's Law
2. Forget about taking this to the next level of scalability. You can not do this on MPPs at the machine wide level.
PGAS with Co-Array Fortran (now Fortran 2008)

Co-array synchronization is at the heart of the typical Co-Array Fortran program. Here is how to exchange an array with your north and south neighbors:

```fortran
COMMON/XCTILB4/ B(N,4)[*]
SAVE /XCTILB4/

CALL SYNC_ALL( WAIT=(/IMG_S,IMG_N/) )
B(:,3) = B(:,1)[IMG_S]
B(:,4) = B(:,2)[IMG_N]
CALL SYNC_ALL( WAIT=(/IMG_S,IMG_N/) )
```

Lots more examples at co-array.org.
Partitioned Global Address Space: (PGAS)

Multiple threads share at least a part of a global address space.

Can access local and remote data with same mechanisms.

Can distinguish between local and remote data with some sort of typing.

Variants:
- Co-Array Fortran (CAF)
- Fortran 2008
- Unified Parallel C (UPC)

Pros:
1. Looks like SMP on a distributed memory machine.
2. Currently translates code into an underlying message passing version for efficiency.

Cons:
1. Depends on (2) to be efficient.
2. Can easily write lots of expensive remote memory access without paying attention.
3. Currently immature.
Frameworks

One of the more experimental approaches that was gaining some traction was to use a parallel framework that handle the load balancing and messaging while you “fill in” the science. Charm++ is the most popular example:

Charm++

- Object-oriented parallel extension to C++
- Run-time engine allows work to be “scheduled” on the computer.
- High-dynamic, extreme load-balancing capabilities.
- Completely asynchronous.
- NAMD, a very popular MD simulation engine is written in Charm++
After a long time with no positive reports in this space, I can definitely say that the Machine Learning (Artificial Intelligence) community has embraced this in an effective manner.

The most popular frameworks/toolkits/packages used for deep learning (aka Neural Nets) are very much in this philosophy.

PyTorch, TensorFlow and others use a high level descriptive approach to arrange other components, often themselves a higher level layer in Python or whatnot, to invoke libraries written in C++ (and actually Fortran is hidden in there more often than those groups would believe in the form of BLAS type libraries).

These frameworks use threads, GPUs and distributed nodes very heavily.

You could say that the math library nature of this work makes this unique, but the innovation in arranging these codeflows is not at all rote.
Some Alternatives

- OpenCL (Khronos Group)
  - Everyone supports, but not as a primary focus
  - Intel – OpenMP
  - NVIDIA – CUDA, OpenACC
  - AMD – now HIP
    - Khronos has now brought out SYCL
- Fortran 2008+ threads (sophisticated but not consistently implemented)
- C++11 threads are basic (no loops) but better than POSIX
- Python threads are fake (due to Global Interpreter Lock)
- DirectCompute (Microsoft) is not HPC oriented
- C++ AMP (MS/AMD)
- TBB (Intel C++ template library)
- Cilk (Intel, now in a gcc branch)
- Intel oneAPI (Includes DPC++ and extends SYCL)
- Kokkos
Lampson's Law

Butler Lampson famously said "All problems in computer science can be solved by another level of indirection."

The oft cited corollary to this is, "...except for the problem of too many layers of indirection."

This has become a common theme with modern, deep, fragile software stacks.

**cmake**: makes makefiles
  to solve problem of different build environments

**Jupyter** notebooks: layer psudo-html and python on javascript
  to solve lack of standard interactive GUI (kind of like Flash did for the Web)

**pip install**: fetches multiple levels of dependencies (even with a `-r requirements.txt` file)
  to solve Python dependency hell
  but it only made it worse so
  Anaconda tries to help
  until something is missing...

**Containers** try to sweep the whole mess under the carpet (and then roll it up)
  until something is missing...

The list goes on....
Parallel computing isn't immune to Lampson's Law either. The failure of OpenACC and OpenMP to merge, coupled with a number of non-standard, industry supported wannabees has led to several (of course) proposed layers on top of the standardized ones...

I mentioned them briefly in our earlier talks, but this is a good time to mention it again. I'll paraphrase heavily from a decent summary podcast that is mentioned here:


Two of those are **Kokkos** and **RAJA**, which are portability layers from national labs. These two sit one level higher in the abstraction over the vendor-specific programming models. Another national lab team is using RAJA to enable their performance portability along with a related project, **Umpire**, for memory management.

RAJA itself is a project that’s focused on loop execution, and it does not include memory management. RAJA applications either can use the underlying backend’s memory management model directly or they use this related project, Umpire. Umpire is another abstraction layer that sits on top of these other programming models, providing a single interface to do memory management such as mem copies and allocations. This design is nice because it offers a separation of concerns. RAJA is very much focused on executing your kernel on the GPU while Umpire is just focused on memory management and moving memory around.

RAJA is being developed to run on GPUs using **DPC++**. That is part of Intel’s **oneAPI**, and it’s an implementation of the **SYCL** programming model. SYCL is an open programming standard that’s focused on heterogeneous C++ programming. DPC++ is kind of an overloaded term. It refers to the name of Intel’s implementation of the SYCL standard, with some important extensions that they made to the language.

If this seems confusing and redundant, you aren’t wrong. Beware any magic bullet that hasn’t proven itself, regardless of the enthusiasm.
Hybrid Coding

- Problem: given the engineering constraint of a machine made up of a large collection of multi-core processors, how do we use message passing at the wide level while still taking advantage of the local shared memory?

- Similar Problem: given a large machine with accelerators on each node (GPU or MIC), how do we exploit this architecture?

- Solution: Hybrid Coding. Technically, this could be any mix of paradigms. Currently, this is likely MPI with a directive based approach mixed in.

- At the node level, you may find OpenMP or OpenACC directives most usable.

- But, **one must design the MPI layer first, and then apply the OpenMP/ACC code at the node level.** The reverse is rarely a viable option.
Counterintuitive: MPI vs. OpenMP on a node

It might seem obvious that, since OpenMP is created to deal with SMP code, you would ideally like to use that at the node level, even if you use MPI for big scalability across an MPP.

Very often, it turns out that the MPI-to-the-core (pun completely intended) version is faster. This indeed seems odd.

The answer is that after going to the trouble of doing a proper MPI data decomposition, you have also greatly aided the caching mechanism (by moving concurrently accessed data into different regions of memory). Hence the win.

However, if you are only interested in node-level scaling, this would be a lot of effort.
Parallel Programming in a Nutshell

Assuming you just took our workshop

- You have to spread *something* out.
- These can theoretically be many types of abstractions: work, threads, tasks, processes, data,…
- But what they *will* be is your data. And then you will use MPI, and possibly OpenMP/ACC, to operate on that data.
Domain Decomposition Done Well: Load Balanced

- A parallel algorithm can only be as fast as the slowest chunk.
  - Might be dynamic (hurricane moving up coast)
- Communication will take time
  - Usually orders of magnitude difference between registers, cache, memory, network/remote memory, disk
  - Data locality and “neighborly-ness” matters very much.

Is Texas vs. New Jersey a good idea?
In Conclusion…

OpenMP

OpenACC

MPI
The Future and where you fit.

While the need is great, there is only a short list of serious contenders for 2020 exascale computing usability.

**MPI 3.0 +X** (MPI 3.0 specifically addresses exascale computing issues)

**PGAS** (partitioned global address space)

CAF (now in Fortran 2008!), UPC

**APGAS**

**X10, Chapel**

Frameworks

**Charm++**

Functional

Haskell
Do you really care about software?

Of course you should. Here are a few reassuring words that software at exascale is not an afterthought, followed by more than a few application examples.
## ECP application domains.

<table>
<thead>
<tr>
<th>National security</th>
<th>Energy security</th>
<th>Economic security</th>
<th>Scientific discovery</th>
<th>Earth system</th>
<th>Health care</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stockpile stewardship</td>
<td>Turbine wind plant efficiency</td>
<td>Additive manufacturing of qualifiable metal parts</td>
<td>Find, predict, and control materials and properties</td>
<td>Accurate regional impact assessments in Earth system models</td>
<td>Accelerate and translate cancer research</td>
</tr>
<tr>
<td>Next generation simulation tools for assessing nuclear weapons performance</td>
<td>High-efficiency, low-emission combustion engine and gas turbine design</td>
<td>Reliable and efficient planning of the power grid</td>
<td>Cosmological probe of the standard model of particle physics</td>
<td>Stress-resistant crop analysis and catalytic conversion of biomass-derived alcohols</td>
<td></td>
</tr>
<tr>
<td>Response to hostile threat environments and reentry conditions</td>
<td>Materials design for extreme environments of nuclear fission and fusion reactors</td>
<td>Seismic hazard risk assessment</td>
<td>Validate fundamental laws of nature</td>
<td>Metagenomics for analysis of biogeochemical cycles, climate change, environmental remediation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Design and commercialization of Small Modular Reactors</td>
<td></td>
<td>Demystify origin of chemical elements</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Subsurface use for carbon capture, petroleum extraction, waste disposal</td>
<td></td>
<td>Light source-enabled analysis of protein and molecular structure and design</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Scale-up of clean fossil fuel combustion</td>
<td></td>
<td>Whole-device model of magnetically confined fusion plasmas</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Biofuel catalyst design</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
xSDK Version 0.4.0: December 2018 (even better today)

https://xsdk.info

Multiphysics Application C

Application A

Application B

xSDK functionality, Dec 2018

Tested on key machines at ALCF, NERSC, OLCF, also Linux, Mac OS X

Each xSDK member package uses or can be used with one or more xSDK packages, and the connecting interface is regularly tested for regressions.

xSDK

More external software

HDF5

BLAS

More libraries

Spack

Extreme-Scale Scientific Software Development Kit (xSDK)

Domain components
- Reacting flow, etc.
- Reusable.

Libraries
- Solvers, etc.
- Interoperable.

Frameworks & tools
- Doc generators.
- Test, build framework.

SW engineering
- Productivity tools.
- Models, processes.

Impact: Improved code quality, usability, access, sustainability

Foundation for work on performance portability, deeper levels of package interoperability

December 2018
- 17 math libraries
- 2 domain components
- 16 mandatory xSDK community policies
- Spack xSDK installer

Alquimia

SLEPc

hypre

SUNDIALS

AMReX

Omega_h

deal.II

PETSc

SuperLU

Trilinos

MFEM

DTK

Tasmanian

MAGMA

PLASMA

HDF5

SuperLU

TRILINOS

SUNDIALS

MFEM

AMReX

Omega_h

deal.II

Spack

Domain components

Libraries

Frameworks & tools

SW engineering

Impact: Improved code quality, usability, access, sustainability

Foundation for work on performance portability, deeper levels of package interoperability
The planned ECP ST SDKs will span all technology areas

<table>
<thead>
<tr>
<th>xSDK (16)</th>
<th>PMR Core (17)</th>
<th>Tools and Technology (11)</th>
<th>Compilers &amp; Support (7)</th>
<th>Visualization Analysis &amp; Reduction (9)</th>
<th>Data Mgmt, I/O Services, &amp; Checkpoint restart (12)</th>
<th>Ecosystem/E4S at-large (12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hypre</td>
<td>Legion</td>
<td>TAU</td>
<td>openarc</td>
<td>ParaView</td>
<td>FAODEL</td>
<td>BEE</td>
</tr>
<tr>
<td>FileSCI</td>
<td>Kokkos (Support)</td>
<td>HPCToolkit</td>
<td>Kitsune</td>
<td>Catalyst</td>
<td>ROMIO</td>
<td>FSEFI</td>
</tr>
<tr>
<td>MFEM</td>
<td>RAJA</td>
<td>Dyninst Binary Tools</td>
<td>LLVM</td>
<td>VTK-m</td>
<td>Mercury (part of Mochi suite)</td>
<td>Kitten Lightweight Kernel</td>
</tr>
<tr>
<td>Kokkoskernels</td>
<td>CHAI</td>
<td>Gotcha</td>
<td>CHILL Autotuning Compiler</td>
<td>SZ</td>
<td>HDF5</td>
<td>COOLR</td>
</tr>
<tr>
<td>Trilinos</td>
<td>PaRSEC*</td>
<td>Caliper</td>
<td>LLVM OpenMP compiler</td>
<td>zfp</td>
<td>Parallel netCDF</td>
<td>NRM</td>
</tr>
<tr>
<td>SUNDIALS</td>
<td>DARMA</td>
<td>Program Database Toolkit</td>
<td>OpenMP V &amp; V</td>
<td>VisIt</td>
<td>ADIOS</td>
<td>ArgoContainers</td>
</tr>
<tr>
<td>PETSc/TAO</td>
<td>GASNet-EX</td>
<td>Search using Random Forests</td>
<td>Flang/LLVM Fortran compiler</td>
<td>ASCENT</td>
<td>Darshan</td>
<td>Spack</td>
</tr>
<tr>
<td>libEnsemble</td>
<td>Qthreads</td>
<td></td>
<td></td>
<td>Cinema</td>
<td>UnityCR</td>
<td>MarFS</td>
</tr>
<tr>
<td>STRUMPACK</td>
<td>BOLT</td>
<td></td>
<td></td>
<td>ROVER</td>
<td>VeloC</td>
<td>GUFI</td>
</tr>
<tr>
<td>SuperLU</td>
<td>UPC++</td>
<td></td>
<td></td>
<td></td>
<td>IOSS</td>
<td>Intel GEOPM</td>
</tr>
<tr>
<td>ForTrilinos</td>
<td>MPICH</td>
<td></td>
<td></td>
<td></td>
<td>HXHIM</td>
<td>mpiFileUtils</td>
</tr>
<tr>
<td>SLATE</td>
<td>Open MPI</td>
<td></td>
<td></td>
<td></td>
<td>SCR</td>
<td>TriBITS</td>
</tr>
<tr>
<td>MAGMA</td>
<td>Umpire</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DTK</td>
<td>QUO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tasmanian</td>
<td>Papyrus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TuckerMPI</td>
<td>SICM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AML</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Key:
- PMR
- Tools
- Math Libraries
- Data and Vis
- Ecosystems and Delivery
Appendix

Endless apps...
CEED is targeting several ECP applications

- Compressible flow (MARBL)
- Climate (E3SM)
- Modular Nuclear Reactors (ExaSMR)
- Wind Energy (ExaWind)
- Additive Manufacturing (ExaAM)
- Combustion (Nek5000)
- Subsurface (GEOS)
- Urban systems (Urban)
- Magnetic Fusion (WDMApp)

PI: Tzanio Kolev (LLNL)
ECP’s Adaptive Mesh Refinement Co-Design Center: AMReX

- Develop and deploy software to support block-structured adaptive mesh refinement on exascale architectures
  - Core AMR functionality
  - Particles coupled to AMR meshes
  - Embedded boundary (EB) representation of complex geometry
  - Linear solvers
  - Supports two modalities of use
    - Library support for AMR
    - Framework for constructing AMR applications

- Provide direct support to ECP applications that need AMR for their application

- Evaluate software technologies and integrate with AMReX when appropriate

- Interact with hardware technologies / vendors

PI: John Bell (LBNL)
Goal: Replace the activities in HPC workflow that have been mediated through file I/O with in-situ methods / workflows. data reduction, analysis, code coupling, aggregation (e.g. parameter studies).

Cross-cutting tools:

- Workflow setup, manager (Cheetah, Savanna); Data coupler (ADIOS-SST); Compression methods (MGARD, FTK, SZ), compression checker (Z-checker)

- Performance tools (TAU, Chimbuco, SOSFlow)
ECP’s Co-Design Center for Particle Applications: CoPA

**Goal**: Develop algorithms and software for particle methods,

**Cross-cutting capabilities:**

- Specialized solvers for quantum molecular dynamics (Progress / BML).
- Performance-portable libraries for classical particle methods in MD, PDE (Cabana).
- FFT-based Poisson solvers for long-range forces.

**Technical approach:**

- High-level C++ APIs, plus a Fortran interface (Cabana).
- Leverage existing / planned FFT software.
- Extensive use of miniapps / proxy apps as part of the development process.

<table>
<thead>
<tr>
<th>Molecular Dynamics (MD)</th>
<th>MD (w/ long-range)</th>
<th>Particle-in-Cell (PIC)</th>
<th>O(N) Quantum</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Construct neighbor lists</td>
<td>2. Construct neighbor lists</td>
<td>2. (Poisson) field solve on mesh</td>
<td>2. Construct neighbor lists</td>
</tr>
<tr>
<td>3. Compute forces on particles due to short-range neighbors HAGGmk kernel</td>
<td>3. Compute forces on particles due to short-range neighbors</td>
<td>3. Gather forces (mesh → particles)</td>
<td>3. Build Hamiltonian (H) and overlap (S) matrices</td>
</tr>
<tr>
<td>4. Integrate equations of motion (particle update)</td>
<td>4. Compute (approximate) long-range forces, e.g. via Ewald, PIM, FMM, or tree methods</td>
<td>4. Particle push (update)</td>
<td>4. Invert S and orthogonalize H</td>
</tr>
<tr>
<td>5. Resorting of particles into cell lists</td>
<td>5. Integrate equations of motion (particle update)</td>
<td>May be done occasionally/never (5. Remapping: generate new set of particles)</td>
<td>5. SCF loop iteration</td>
</tr>
</tbody>
</table>

**Communication**

- May not be done every timestep, or never if using cell lists for simple bonded interactions

**Memory/flop kernel**

- 1. Halo exchange of ghost cells
- 2. Construct neighbor lists
- 3. Build Hamiltonian (H) and overlap (S) matrices

**O(N) Quantum**

- Calculate partial charges via Coulomb sum
- Update H and, if necessary, orthogonalize
- Compute forces on atoms

---

PI: Sue Mniszewski (LANL) recently replacing Tim Germann (LANL), who is taking on a larger role in ECP.
ECP’s Co-Design Center for Machine Learning: ExaLearn
Bringing together experts from 8 DOE Laboratories

- AI has the potential to accelerate scientific discovery or enable prediction in areas currently too complex for direct simulation (ML for HPC and HPC for ML)

- AI use cases of interest to ECP:
  - *Classification and regression*, including but not limited to image classification and analysis, e.g. scientific data output from DOE experimental facilities or from national security programs.
  - *Surrogate models* in high-fidelity and multiscale simulations, including uncertainty quantification and error estimation.
  - *Structure-to-function relationships*, including genome-to-phenome, the prediction of materials performance based on atomistic structures, or the prediction of performance margins based on manufacturing data.
  - *Control systems*, e.g., for wind plants, nuclear power plants, experimental steering and autonomous vehicles.
  - *Inverse problems* and optimization. This area would include, for example, inverse imaging and materials design.

- Areas in need of research
  - Data quality and statistics
  - Learning algorithms
  - Physics-Informed AI
  - Verification and Validation
  - Performance and scalability
  - Workflow and deployment

**Expected Work Product: A Toolset That . . .**

- Has a line-of-sight to exascale computing, e.g. through using exascale platforms directly, or providing essential components to an exascale workflow
- Does not replicate capabilities easily obtainable from existing, widely-available packages
- Builds in domain knowledge where possible “Physics”-based ML and AI
- Quantifies uncertainty in predictive capacity
- Is interpretable
- Is reproducible
- Tracks provenance

PI: Frank Alexander (BNL)
Machine Learning in the Light Source Workflow

**Beam Line Control and Data Acquisition (DAQ)**

- Data TB/s
- ML to design light source beam lines
- ML at DAQ to control data as it is acquired
- ML to control the beam line parameters

**Local Systems**

- Compressor Nodes
- Online Monitoring and Fast Feedback
- Data
- ML for data compression (e.g., hit finding). Use models learned remotely.
- ML for fast analysis at the experimental facility. Uses models learned remotely.

**Network**

- 10 GB/s - 1Tb/s
- Model
- ML networks for image classification, feature detection and solving inverse problems (how to change experiment params to get desired experiment result)

**Remote Exascale HPC**

- Exascale Supercomputer
- Data
- Model
- Model
- Simulate experiments, beam line control and diffraction images at scale to create data for training

PI: Frank Alexander (BNL)
Exascale apps can deliver transformative products and solutions

<table>
<thead>
<tr>
<th>ExaWind</th>
<th>ExaAM</th>
<th>EQSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Turbine Wind Plant Efficiency</strong> <em>(Mike Sprague, NREL)</em>&lt;br&gt;• Harden wind plant design and layout against energy loss susceptibility&lt;br&gt;• Increase penetration of wind energy&lt;br&gt;<em>Challenges</em>: linear solver perf in strong scale limit; manipulation of large meshes; overset of structured &amp; unstructured grids; communication-avoiding linear solvers</td>
<td><strong>Additive Manufacturing (AM) of Qualifiable Metal Parts</strong> <em>(John Turner, ORNL)</em>&lt;br&gt;• Accelerate the widespread adoption of AM by enabling routine fabrication of qualifiable metal parts&lt;br&gt;<em>Challenges</em>: capturing unresolved physics; multi-grid linear solver performance; coupled physics</td>
<td><strong>Earthquake Hazard Risk Assessment</strong> <em>(David McCallen, LBNL)</em>&lt;br&gt;• Replace conservative and costly earthquake retrofits with safe purpose-fit retrofits and designs&lt;br&gt;<em>Challenges</em>: full waveform inversion algorithms</td>
</tr>
</tbody>
</table>
EIQSIM: Understanding and predicting earthquake phenomenon

- Vertical motion
- Horizontal motion
- Site ground motions
- Surface waves
- Body waves
- Site
- Path
- Source
- Ground motions tend to be very site specific

PI: David McCallen (LBNL)
EQSIM: The Exascale “Big Lift” – regional ground motion simulations at engineering frequencies

Pipelines     Long-span Bridges     Tall Buildings     Low-rise Buildings and Industrial Facilities     Energy System Components     Nuclear Power Equipment

Exascale objective

0. Hz  0.1 Hz  0.2 Hz  1.0 Hz  2.0 Hz  3.0 Hz  10.0 Hz  25.0 Hz

Range of historical ground motion simulations

Doubling the frequency resolution = 16X computational effort!

PI: David McCallen (LBNL)
Exascale apps can deliver transformative products and solutions

MFIX-Exa

Scale-up of Clean Fossil Fuel Combustion
(Madhava Syamlal, NETL)
• Commercial-scale demonstration of transformational energy technologies – curbing CO₂ emissions at fossil fuel power plants by 2030

Challenges: load balancing; strong scaling thru transients

GAMESS

Biofuel Catalyst Design
(Mark Gordon, Ames)
• Design more robust and selective catalysts orders of magnitude more efficient at temperatures hundreds of degrees lower

Challenges: weak scaling of overall problem; on-node performance of molecular fragments

EXAALT

Materials for Extreme Environments
(Danny Perez, LANL)
• Simultaneously address time, length, and accuracy requirements for predictive microstructural evolution of materials

Challenges: SNAP kernel efficiency on accelerators; efficiency of DFTB application on accelerators
Exascale apps can deliver transformative products and solutions

<table>
<thead>
<tr>
<th>ExaSMR</th>
</tr>
</thead>
</table>
| **Design and Commercialization of Small Modular Reactors**  
(Steve Hamilton, ORNL) |
| - Virtual test reactor for advanced designs via experimental-quality simulations of reactor behavior |

*Challenges:* existing GPU-based MC algorithms require rework for hardware less performant for latency-bound algorithms with thread divergence; performance portability with OCCA & OpenACC not achievable; insufficient node memory for adequate CFD + MC coupling

<table>
<thead>
<tr>
<th>Subsurface</th>
</tr>
</thead>
</table>
| **Carbon Capture, Fossil Fuel Extraction, Waste Disposal**  
(Carl Steefel, LBNL) |
| - Reliably guide safe long-term consequential decisions about storage, sequestration, and exploration |

*Challenges:* performance of Lagrangian geomechanics; adequacy of Lagrangian crack mechanics) + Eulerian (reaction, advection, diffusion) models; parallel HDF5 for coupling

<table>
<thead>
<tr>
<th>QMCPACK</th>
</tr>
</thead>
</table>
| **Materials for Extreme Environments**  
(Paul Kent, ORNL) |
| - Find, predict and control materials and properties at the quantum level with unprecedented accuracy for the design novel materials that rely on metal to insulator transitions for high performance electronics, sensing, storage |

*Challenges:* minimizing on-node memory usage; parallel on-node performance of Markov-chain Monte Carlo
Exascale apps can deliver transformative products and solutions

**ExaSGD**

*Reliable and Efficient Planning of the Power Grid*  
(Henry Huang, PNNL)

- Optimize power grid planning, operation, control and improve reliability and efficiency

*Challenges:* parallel performance of nonlinear optimization based on discrete algebraic equations and possible mixed-integer programming

**Combustion-PELE**

*High-Efficiency, Low-Emission Combustion Engine Design*  
(Jackie Chen, SNL)

- Reduce or eliminate current cut-and-try approaches for combustion system design

*Challenges:* performance of chemistry ODE integration on accelerated architectures; linear solver performance for low-Mach algorithm; explicit LES/DNS algorithm not stable
### Accurate Regional Impact Assessment in Earth Systems (Mark Taylor, SNL)

- Forecast water resources and severe weather with increased confidence; address food supply changes

**Challenges:** MMF approach for cloud-resolving model has large biases; adequacy of Fortran MPI+OpenMP for some architectures; Support for OpenMP and OpenACC

### Catalytic Conversion of Biomass-Derived Alcohols (Thom Dunning, PNNL)

- Develop new optimal catalysts while changing the current design processes that remain costly, time consuming, and dominated by trial-and-error

**Challenges:** computation of energy gradients for coupled-cluster implementation; on- and off-node performance

### Exascale apps can deliver transformative products and solutions

### E3SM-MMF

<table>
<thead>
<tr>
<th><strong>Accurate Regional Impact Assessment in Earth Systems</strong> (Mark Taylor, SNL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Forecast water resources and severe weather with increased confidence; address food supply changes</td>
</tr>
<tr>
<td><strong>Challenges:</strong> MMF approach for cloud-resolving model has large biases; adequacy of Fortran MPI+OpenMP for some architectures; Support for OpenMP and OpenACC</td>
</tr>
</tbody>
</table>

### NWChemEx

<table>
<thead>
<tr>
<th><strong>Catalytic Conversion of Biomass-Derived Alcohols</strong> (Thom Dunning, PNNL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Develop new optimal catalysts while changing the current design processes that remain costly, time consuming, and dominated by trial-and-error</td>
</tr>
<tr>
<td><strong>Challenges:</strong> computation of energy gradients for coupled-cluster implementation; on- and off-node performance</td>
</tr>
</tbody>
</table>

### ExaBiome

<table>
<thead>
<tr>
<th><strong>Metagenomics for Analysis of Biogeochemical Cycles</strong> (Kathy Yelick, LBNL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Discover knowledge useful for environmental remediation and the manufacture of novel chemicals and medicines</td>
</tr>
<tr>
<td><strong>Challenges:</strong> Inability of message injection rates to keep up with core counts; efficient and performant implementation of UPC, UPC++, GASNet; GPU performance; I/O performance</td>
</tr>
</tbody>
</table>
E3SM-Multiscale Modeling Framework (MMF)
Cloud Resolving Climate Model for E3SM

- Develop capability to assess regional impacts of climate change on the water cycle that directly affect the US economy such as agriculture and energy production.

- Cloud resolving climate model is needed to reduce major systematic errors in climate simulations due to structural uncertainty in numerical treatments of convection – such as convective storm systems

- Challenge: cloud resolving climate model using traditional approaches requires *zettascale* resources

- E3SM “conventional” approach:
  - Run the E3SM model with a global cloud resolving atmosphere and eddy resolving ocean.
    - 3 km atmosphere/land (7B grid points) and 15-5 km ocean/ice (1B grid points)
  - Achieve throughput rate of 5 SYPD to perform climate simulation campaigns including a 500 year control simulation
  - Detailed benchmarks on KNL and v100 GPUs show negligible speedups compared to conventional CPUs
    - Low arithmetic intensity of most of the code; throughput requirements lead to strong scaling and low work per node.

- E3SM-MMF: Use a multiscale approach ideal for new architectures to achieve cloud resolving convection on Exascale
  - Exascale will make “conventional” cloud resolving simulations routine for shorter simulations (process studies, weather prediction)

For cloud resolving climate simulations, we need fundamentally new approaches to take advantage of exascaleresources

Convective storm system nearing the Chicago metropolitan area
http://www.spc.noaa.gov/misc/AbtDerechos/derechofacts.htm

PI: Mark Taylor (SNL)
Exascale apps can deliver transformative products and solutions

**ExaSky**

**Cosmological Probe of the Standard Model of Particle Physics**
(Salman Habib, ANL)

- Unravel key unknowns in the dynamics of the Universe: dark energy, dark matter, and inflation

*Challenges:* subgrid model accuracy; OpenMP performance on GPUs; file system stability and availability

**LatticeQCD**

**Validate Fundamental Laws of Nature**
(Andreas Kronfeld, FNAL)

- Correct light quark masses; properties of light nuclei from first principles; <1% uncertainty in simple quantities

*Challenges:* performance of critical slowing down; reducing network traffic to reduce system interconnect contention; strong scaling performance to mitigate reliance on checkpointing

**WarpX**

**Plasma Wakefield Accelerator Design**
(Jean-Luc Vay, LBNL)

- Virtual design of 100-stage 1 TeV collider; dramatically cut accelerator size and design cost

*Challenges:* scaling of Maxwell FFT-based solver; maintaining efficiency of large timestep algorithm; load balancing
## Exascale apps can deliver transformative products and solutions

<table>
<thead>
<tr>
<th>WDMApp</th>
<th>ExaStar</th>
<th>ExaFEL</th>
<th>CANDLE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High-Fidelity Whole Device Modeling</strong></td>
<td><strong>Demystify Origin of Chemical Elements</strong></td>
<td><strong>Light Source-Enabled Analysis of Protein and Molecular Structure</strong></td>
<td><strong>Accelerate and Translate Cancer Research</strong></td>
</tr>
<tr>
<td>of Magnetically Confined Fusion Plasmas</td>
<td><em>(Dan Kasen, LBNL)</em></td>
<td><em>(Amadeo Perazzo, SLAC)</em></td>
<td><em>(Rick Stevens, ANL)</em></td>
</tr>
<tr>
<td><em>(Amitava Bhattacharjee, PPPL)</em></td>
<td>• What is the origin of the elements?</td>
<td>• Process data without beam time loss</td>
<td>• Develop predictive preclinical models and accelerate diagnostic and</td>
</tr>
<tr>
<td></td>
<td>• How does matter behave at extreme densities?</td>
<td>• Determine nanoparticle size and shape changes</td>
<td>targeted therapy through predicting mechanisms of RAS/RAF driven cancers</td>
</tr>
<tr>
<td></td>
<td>• What are the sources of gravity waves?</td>
<td>• Engineer functional properties in biology and materials science</td>
<td><strong>Challenges:</strong> increasing accelerator utilization for model search; effectively exploiting HP16; preparing for any data management or communication bottlenecks</td>
</tr>
<tr>
<td></td>
<td><strong>Challenges:</strong> delivering performance on accelerators; delivering fidelity for general relativity implementation</td>
<td><strong>Challenges:</strong> improving the strong scaling (one event processed over many cores) of compute-intensive algorithms (ray tracing, M-TIP) on accelerators</td>
<td></td>
</tr>
</tbody>
</table>