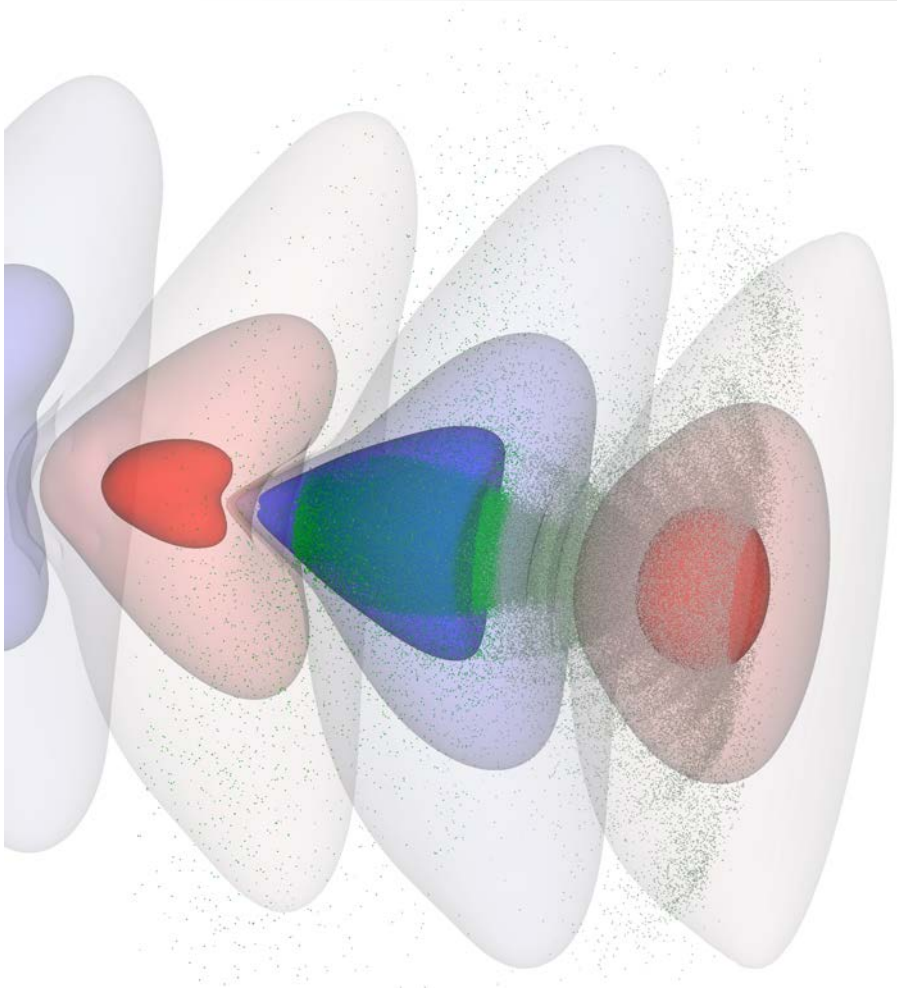
Transformative Petascale Particle-in-Cell Simulation of Laser Plasma Interactions on Blue Waters



Frank Tsung
Viktor K. Decyk
Weiming An
Xinlu Xu
Han Wen

Warren Mori (PI)

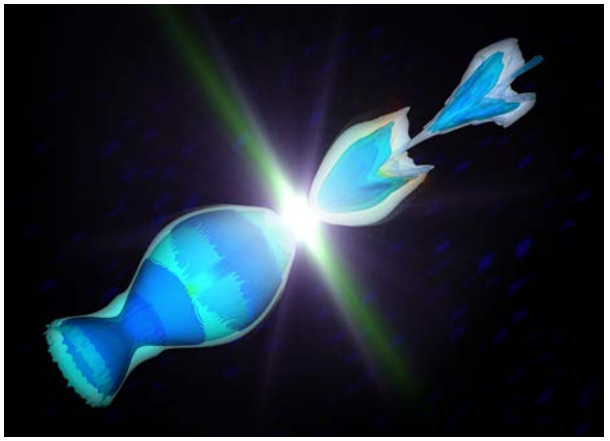
Special Thanks to Galen Arnold & BW Consultants

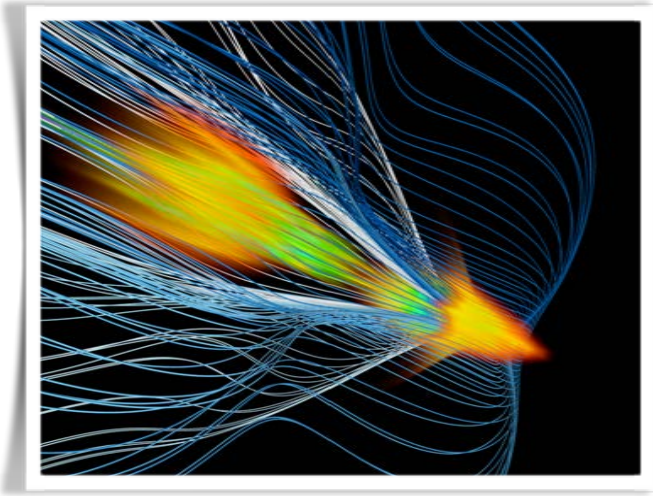
PICKSC UCLA

Summary and Outline

OUTLINE/SUMMARY

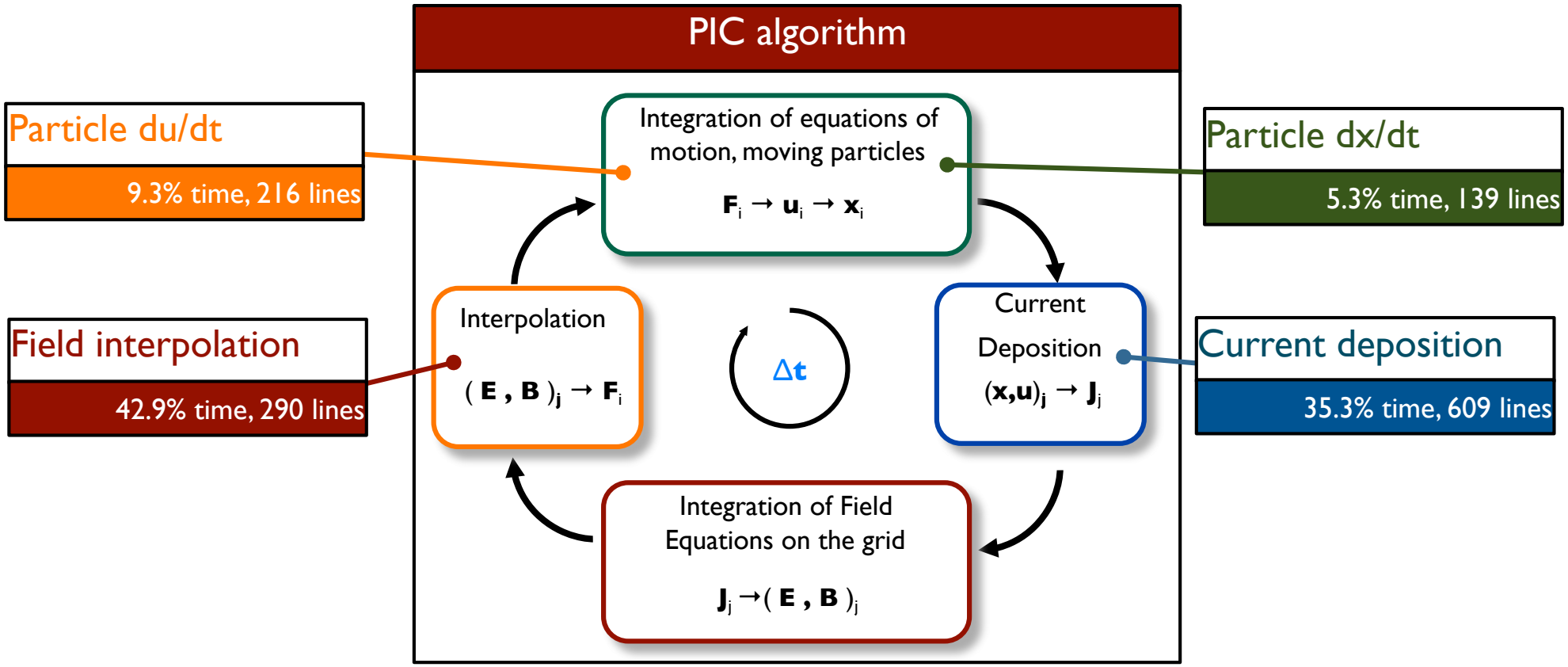
- Overview of the project
  - Particle-in-cell method
  - Our main production code — OSIRIS
- Application of OSIRIS to plasma based accelerators:
  - Production of high quality electrons for ultra-bright XFEL light sources useful for nuclear science (Xu *et al*, being prepared for *Nature Physics* (2017)).
- Higher (2 & 3) dimension simulations of LPI's relevant to laser fusion
  - Adding realisms in 2D LPI simulations relevant to laser fusion.
  - Controlling LPI's by temporal incoherence.
  - Estimates of large scale 3D LPI simulations (& justify the need for new new algorithms on exascale supercomputers)
- Code Developments to reduce simulation time and to move toward exa-scale.
  - Code developments toward exa-scale (multi-GPU's and OpenMP/MPI PIC codes)





UCLA

Profile of OSIRIS + Introduction to PIC



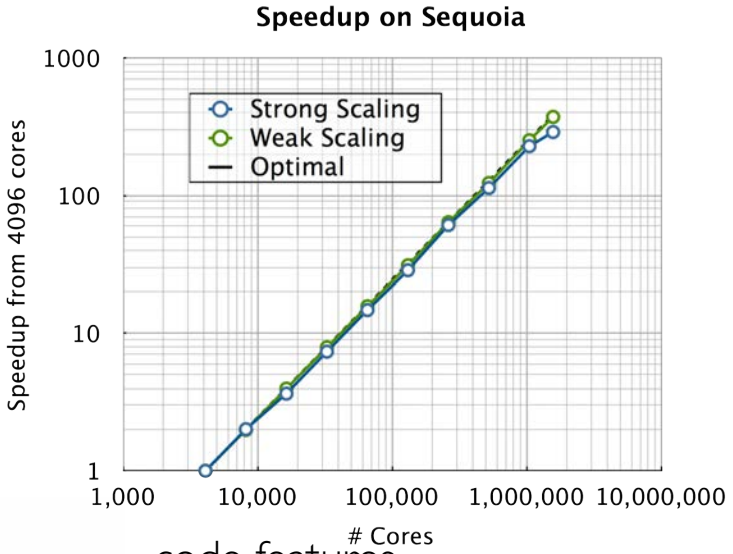
- The particle-in-cell method treats plasma as a collection of computer particles. The interactions does not scale as N^2 due to the fact the particle quantities are deposited on a grids and the interactions are calculated on the grids only. Because (# of particles) >> (# of grids), the timing is dominated by the particle calculations and scales as N (orbit calculation + current & charge deposition).
- The code spends over 90 % of execution time in only 4 routines
- These routines correspond to less than 2 % of the code, optimization and porting is fairly straightforward, although not always trivial.

UCLA

Osiris 3.0

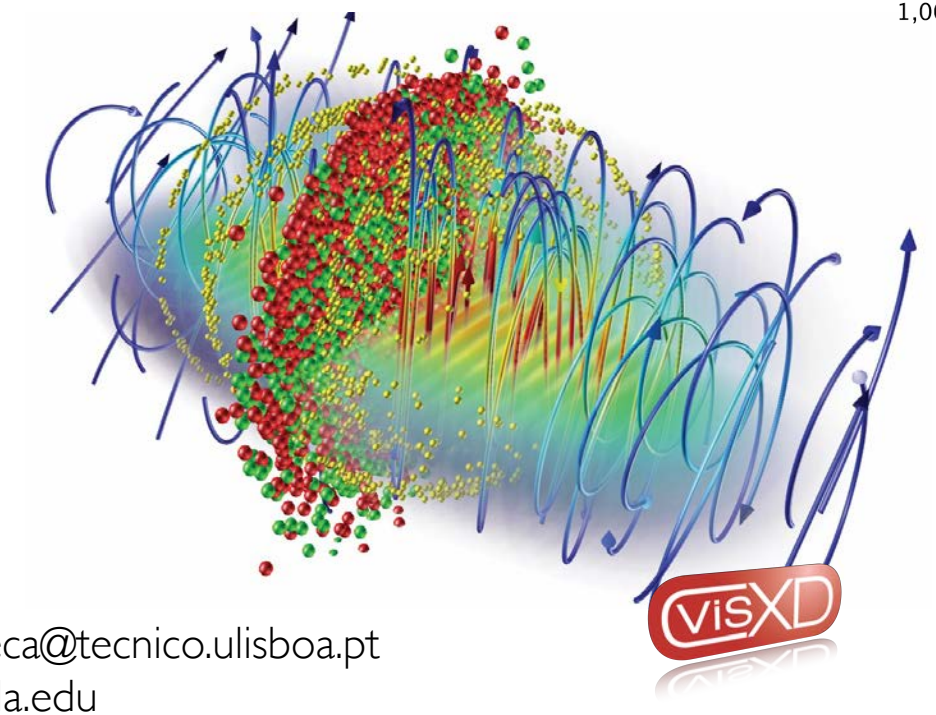
osiris framework

- Massivelly Parallel, Fully Relativistic Particle-in-Cell (PIC) Code
- Visualization and Data Analysis Infrastructure
- Developed by the osiris.consortium
  ⇒ UCLA + IST



TÉCNICO LISBOA

UCLA



code features

- Scalability to ~ 1.6 M cores (on sequoia) and achieved sustained speed of > 2.2PetaFLOPS on Blue Waters
- SIMD hardware optimized
- Parallel I/O
- Dynamic Load Balancing
- QED module
- Particle merging
- OpenMP/MPI parallelism
- CUDA/Intel Phi branches

Ricardo Fonseca: ricardo.fonseca@tecnico.ulisboa.pt
Frank Tsung: tsung@physics.ucla.edu

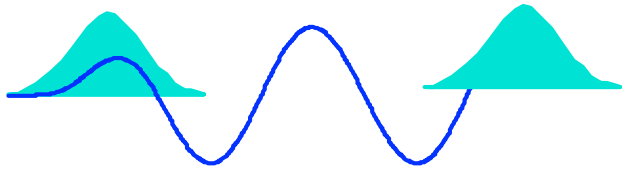
http://epp.tecnico.ulisboa.pt/
http://plasmasim.physics.ucla.edu/

Livingston Curve for Accelerators --- Why plasmas?

UCLA

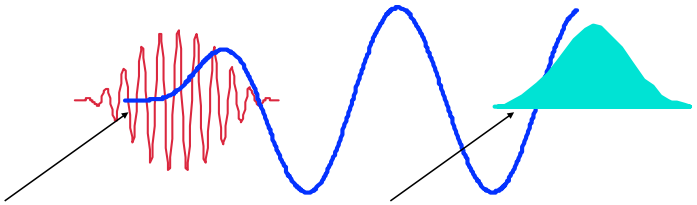
Plasma Wake Field Accelerator(PWFA)

A high energy electron bunch



Laser Wake Field Accelerator(LWFA, SMLWFA)

A single short-pulse of photons

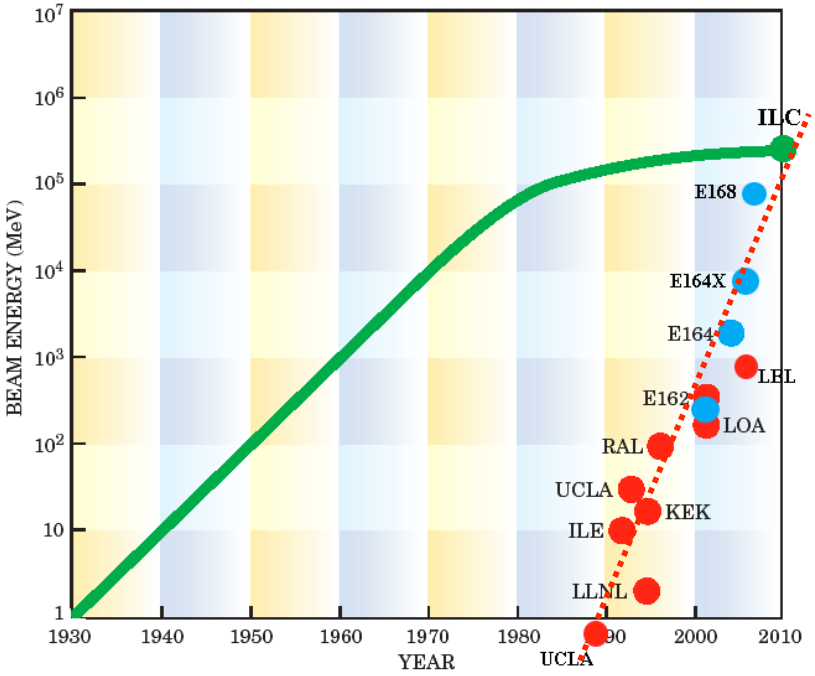


Drive beam

Trailing beam

The Livingston curve traces the history of electron accelerators from Lawrence's cyclotron to present day technology.

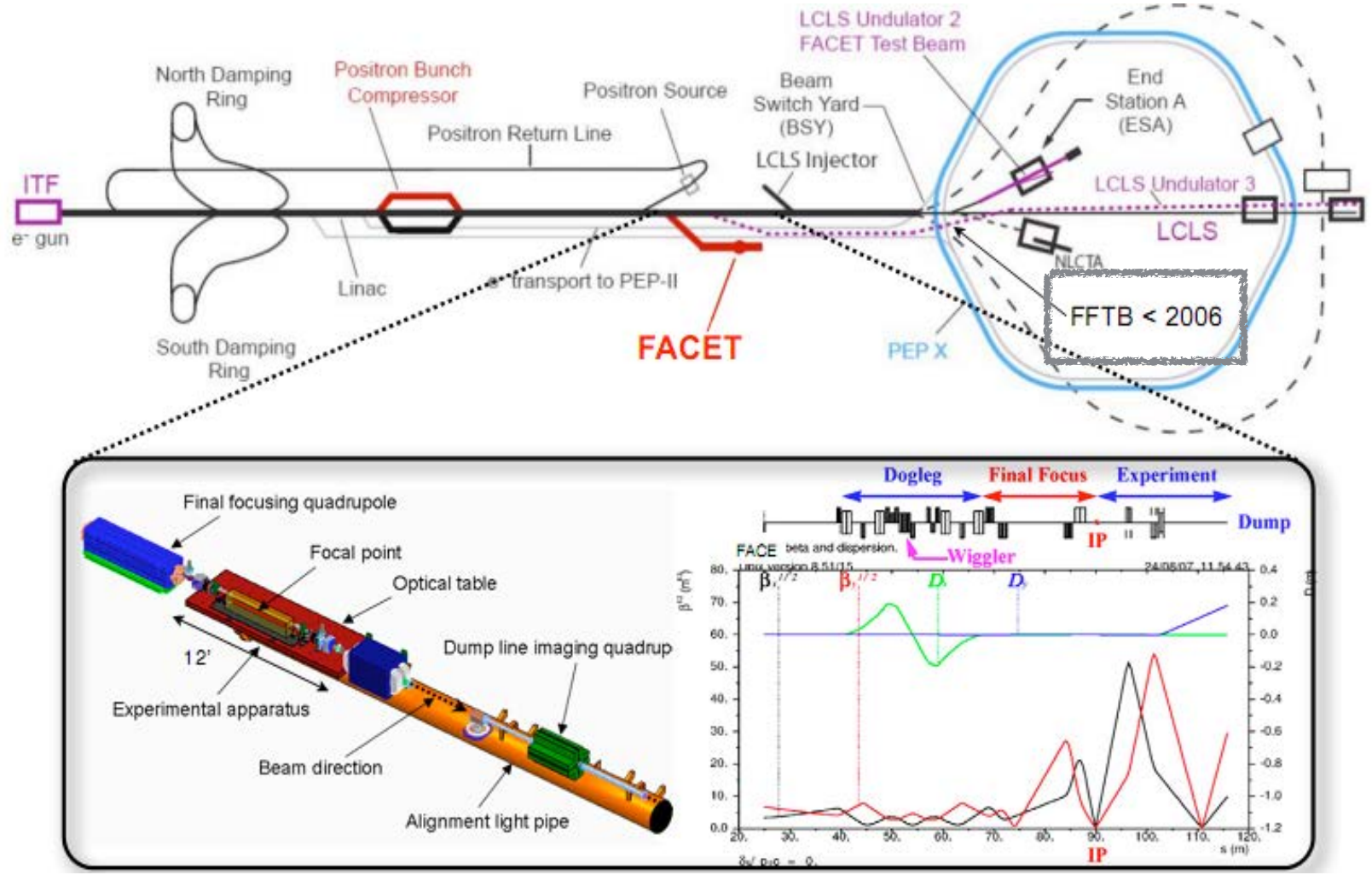
When energies from plasma based accelerators are plotted in the same curve, it shows the exciting trend that within a few years it is will surpass conventional accelerators in terms of energy.



UCLA

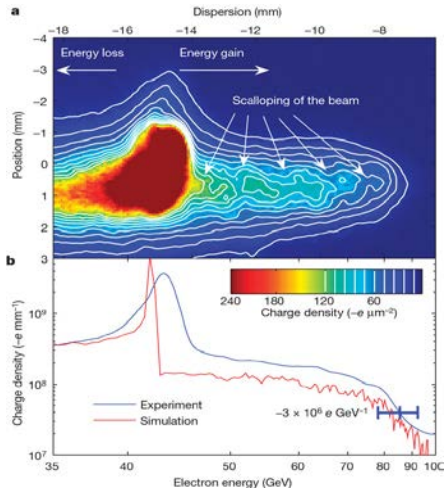
FACET— Plasma based accelerator experiments @ LCLS (linac coherent light source)

Facility for Advanced aCcelerator Experimental Tests

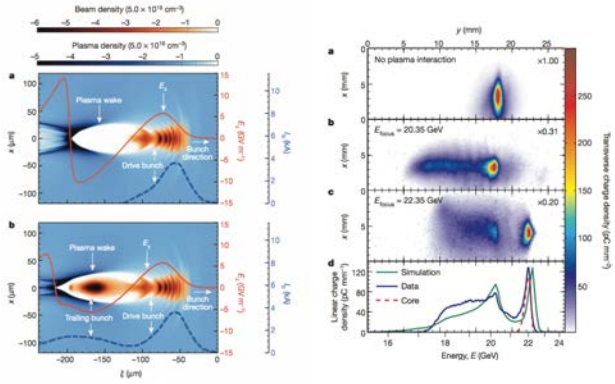


FACET is a new facility to provide high-energy, high peak current e^- & e^+ beams for PWFA experiments at SLAC.

Recent Highlights (in *Nature* journals) in Plasma Based Acceleration (< Last 10 years) -- Simulations play a big role in all of these discoveries!!!

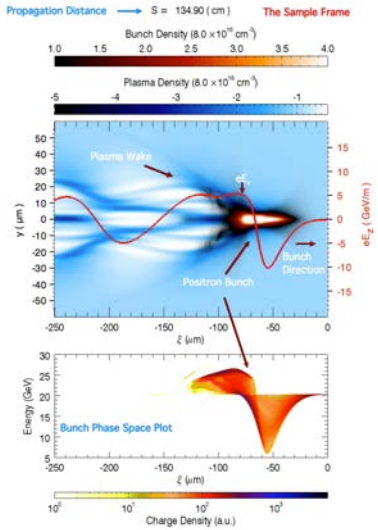


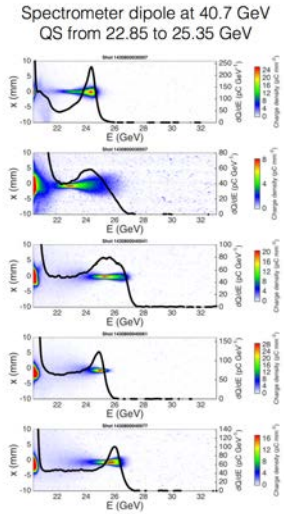
42 GeV in less than one meter! (i.e., 0-42 GeV in 3km, 42-85 GeV in 1m) Simulations also identified ionization induced erosion as the limiting mechanism for energy gain (Blumenfeld et al, (2007)).



2014 "Full Speed Ahead" Cover on *Nature*

***2GeV energy gain in 36 cm of plasma with narrow energy spread. PIC simulations explained the narrow energy spread and produced quantitative agreements between simulation and experiment. (Litos* et al (2014))**





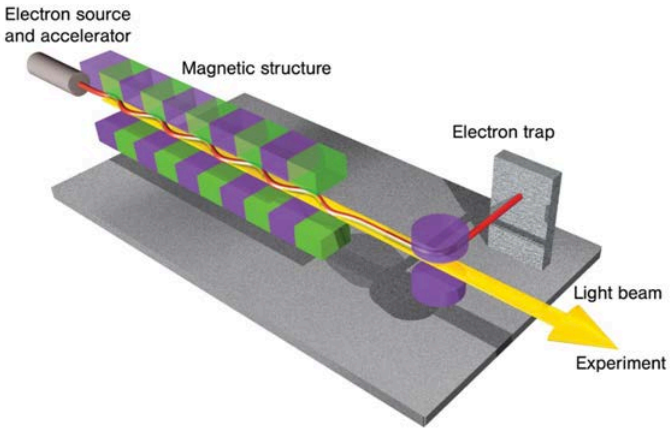
2015 Positron Experiments:

***in 1.3 meters of plasma, positrons gained 3-10GeVs of energy with low energy spread (1.8 to 6%). PIC simulations show that the trailing bunch flattens the wake and produces a high quality positron beam. (Corde* et al (2015)).**

UCLA

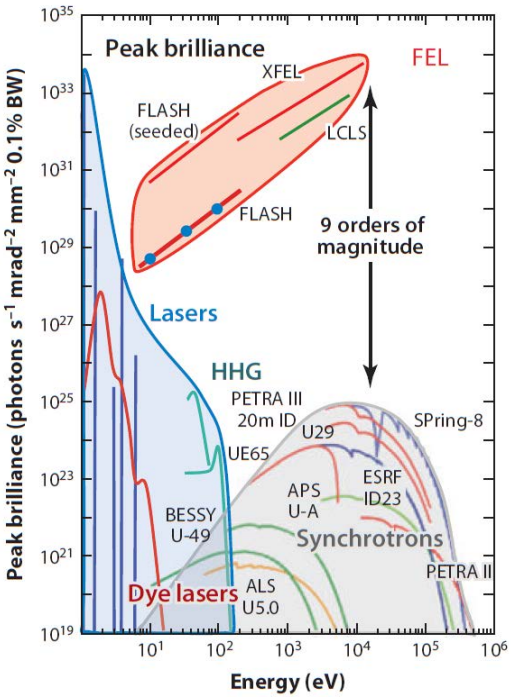
X-ray FEL — Coherent light source at Angstrom scale — Can we make compact radiation sources for nuclear science?

In an X-ray FEL (XFEL), a "coherent" electron beam enters an undulatory and a bright x-ray comes out, the electron beam can be diverted via an magnet (see right).



The need for XFEL's light sources can be justified by looking at the light sources in terms of photon energy and "brilliance". Brilliance, also called brightness, is a measure of the coherence of the photon beam (or roughly the # of photons per volume, where the "time' corresponds the longitudinal distance divided by the speed of light). Improved longitudinal coherence will further increase the brilliance. Shortening thee pulse also produces probes with better time resolution (on the femto-second scale)).

Compared to synchrotron sources, LCLS, which began in 2009, represents a 9 order of magnitude jump in brightness. XFEL's for the first time allow us to probe materials on the nuclear (Angstrom) length scale with femto-second resolution. Laser, while provides high peak brilliance, operates in the ~micron range, which cannot resolve effects on the the nuclear length scale



Using PIC simulations, we are trying to study ways to generate high qualities electron beams with high energy and high quality to produce 20keV (0.62 Angstrom wavelength) lights comparable to those generated at LCLS. XFEL's produced by these beams will allow to probe matter on the nuclear scale. (Bohr radius ~ 1 Angstrom and femto-second resolution).

$$\lambda_r = \frac{\lambda_u}{2\gamma^2}(1 + K^2)$$

(laser wavelength) (beam energy)

UCLA

Question: Is it possible to accelerate a beam with 10^{-4} energy spread using PWFA?

In order for the electrons to radiate coherently in an XFEL, the electrons within a "slice" (i.e., a small neighborhood in the longitudinal direction) needs to be very coherent. Theory estimates that relative energy spread of 10^{-4} is needed for to produce coherent x-ray in the undulator.

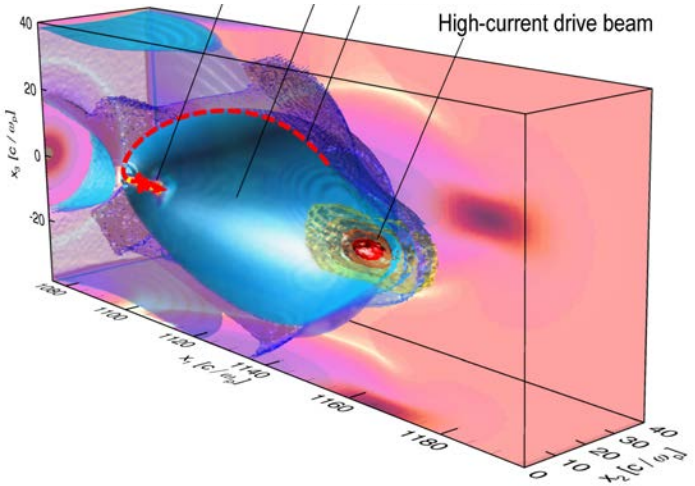
Main Parameters

	I [kA]	σ_z [μm]	ε_n [μm]	σ_r [μm]	Q [pC]	E_b [GeV]	σ_{Eb} [keV]
Driver	2	16	1.2	0.52	269	8	80
	I [kA]	σ_z [μm]	ε_n [μm]	σ_r [μm]	Q [pC]	E_b [GeV]	σ_{Eb} [keV]
Witness	2	6	0.4	0.3	103	8	80
	n_p [cm^{-3}]	k_p^{-1} [μm]					
Plasma	7E+16	20					

QuickPIC simulation setup:

- Box: 167 μm × 167 μm × 167 μm
- Grid Size: 40 nm × 40 nm × 163 nm

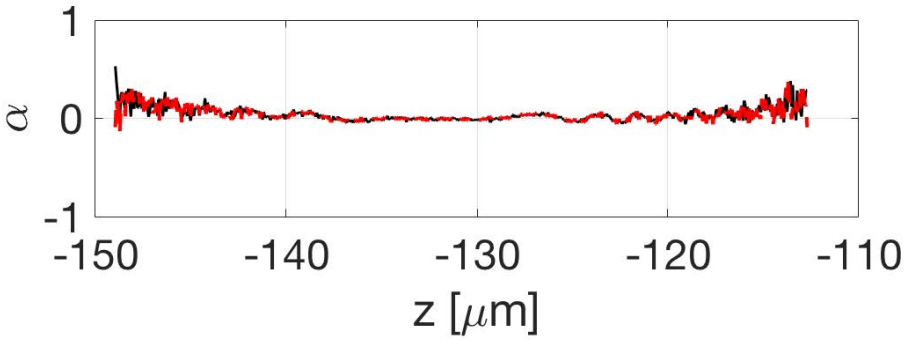
witness beam

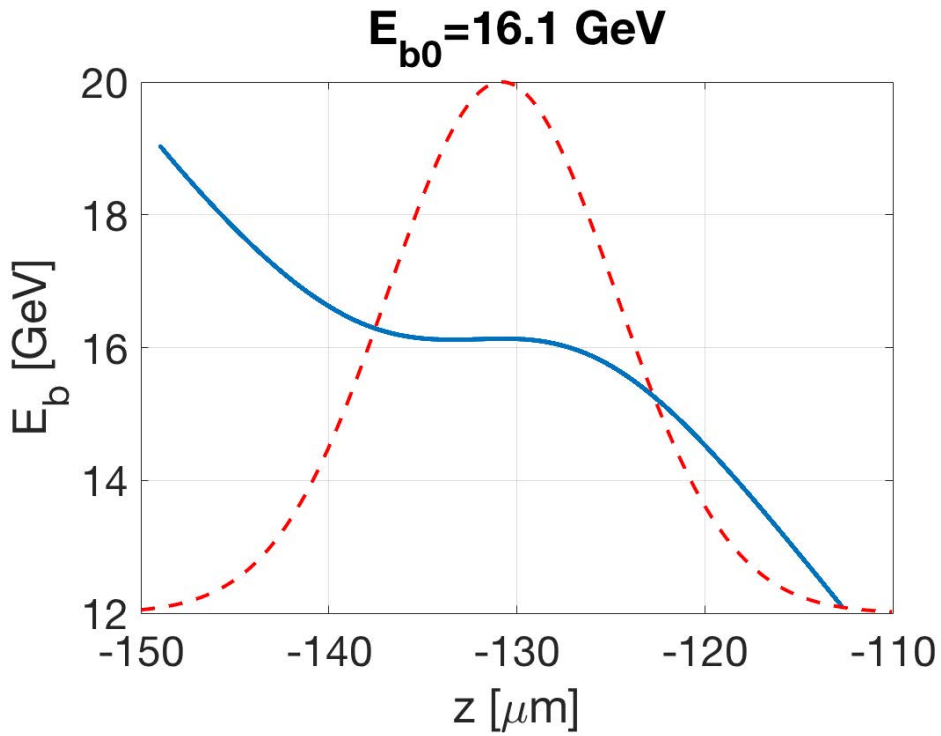


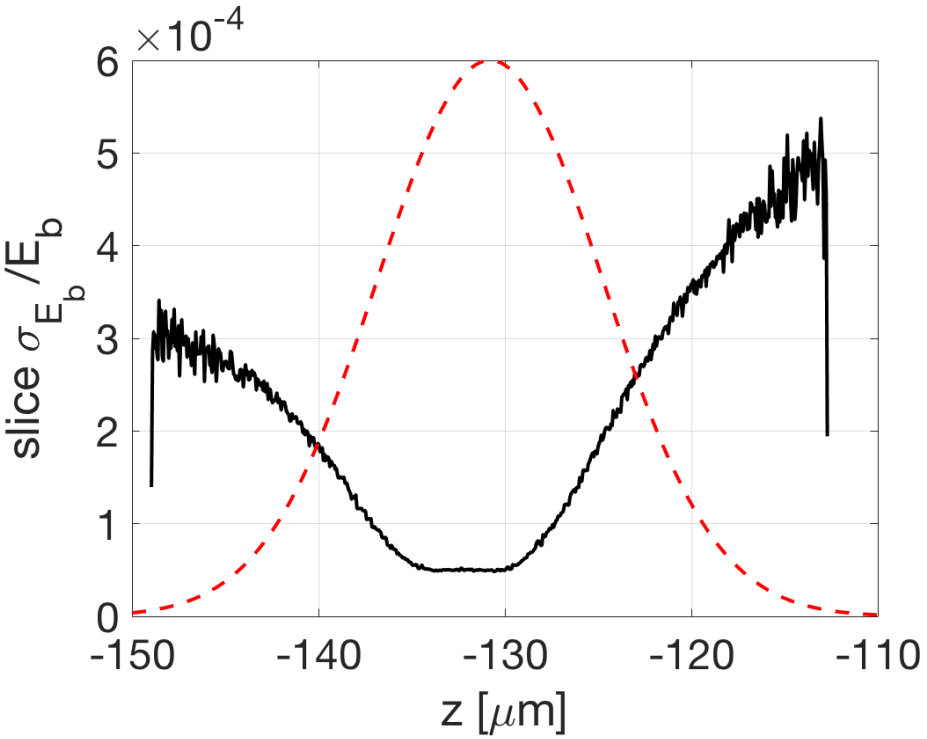
Because the beams are tightly focused and the need to study the beam evolution within a "slice", high resolutions are required in these simulations. Blue Waters resources is critical to study this problem

- 4000 x 4000 x 1000 (16 billion grids)
- 4 million particles in each bunch (642 million real electrons in the witness bunch)
- 64 billion plasma electrons (fixed ions) to 128 billion total plasma particles(with mobile ions)
- 1/2 million core hour / simulation

Simulation results at z=1.55 m







1. "slice" energy in most parts of the beam is less than the requirement: 3×10^{-4}.
2. The emittance (which is the measure of transverse coherence of the beam) is conserved after 1.5 meters.

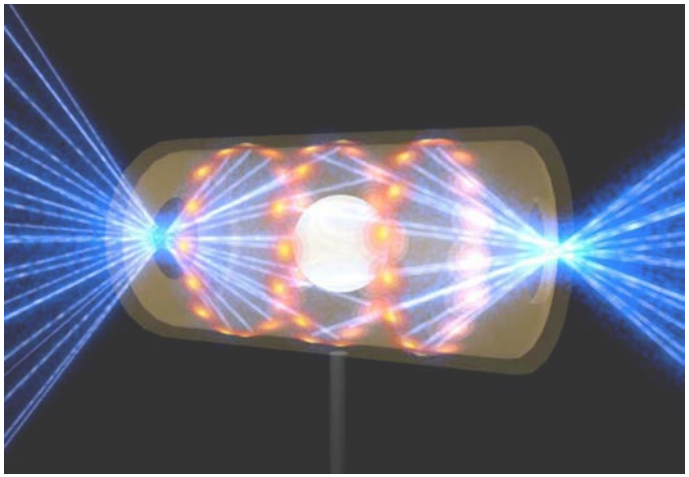
This work is being prepared for publication in *Nature Physics* (X. Xu and co-workers).

Laser Plasma Interactions in IFE

- IFE (inertial fusion energy) uses lasers to compress fusion pellets to fusion conditions. The goal of these experiments is to extract more fusion energy from the fuel than the input energy of the laser. In this case, the excitation of plasma waves via LPI (laser plasma interactions) is detrimental to the experiment in 2 ways.
  - Laser light can be scattered backward toward the source and cannot reach the target
  - LPI produces hot electrons which heats the target, making it harder to compress.

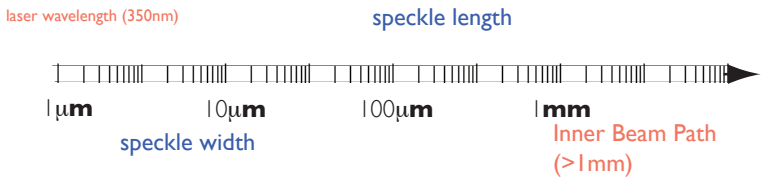
- The LPI problem is very challenging because it spans many orders of magnitude in lengthscale & lengthscale
  - The spatial scale spans from < 1 micron (which is the laser wavelength) to milli-meters (which is the length of the plasma).
  - The temporal scale spans from a femto-second(which is the laser period) to nano-seconds (which is the duration of the fusion pulse). A typical PIC simulation spans ~10ps.

Laser Plasma Interactions

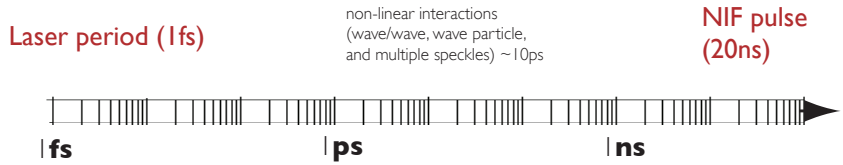


NIF
National Ignition Facility

Lengthscales



Timescales



LPI growth time

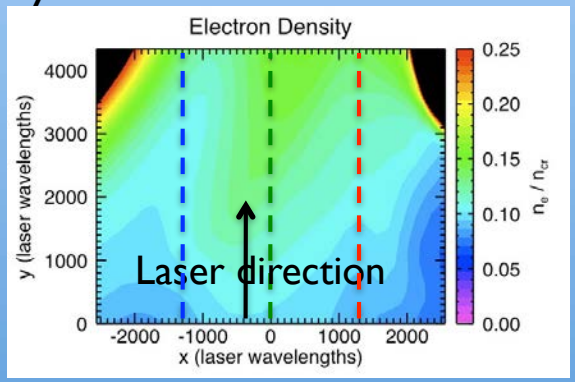
Final laser spike (1ns)

UCLA

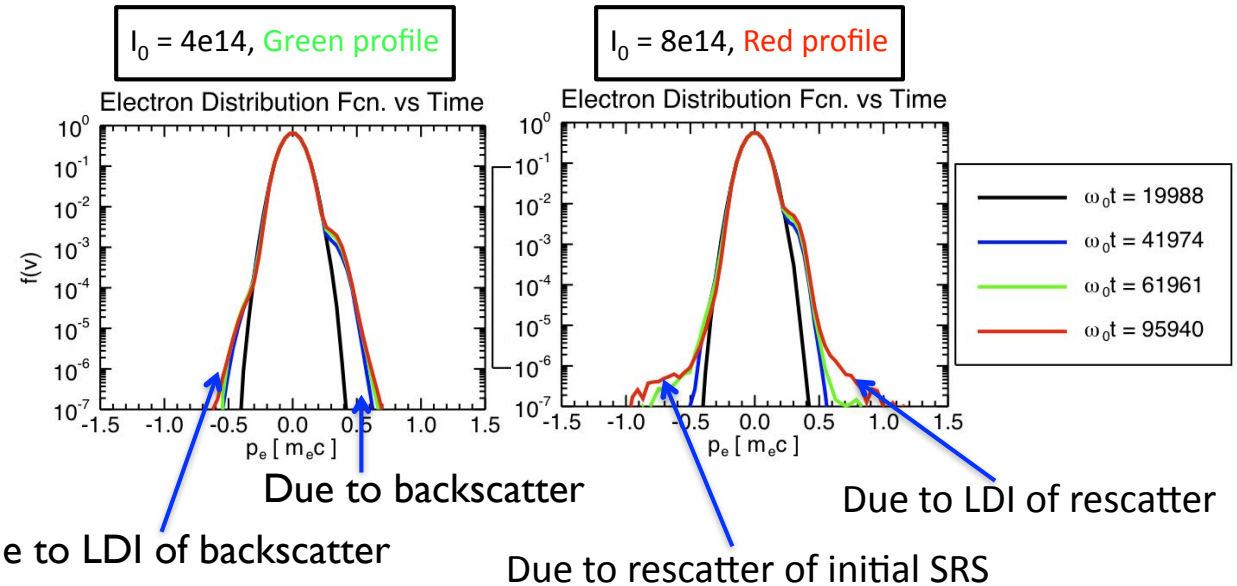
Performing a typical 1D OSIRIS Simulation Using NIF parameters

- Currently, experimentalists @ NIF can re-construct plasma conditions (such as density and temperature) using a hydro code. Using this "plasma map", we can perform a series of 1D OSIRIS simulations along each of the "beam path" indicated by the dash line, each taking ~100 CPU hours.
- 1D OSIRIS Simulations can predict:
  - Spectrum of backscattered lights (which can be compared against experiments)
  - spectrum of energetic electrons (shown below)
  - energy partition, i.e., how the incident laser energy is converted to
    - transmitted light
    - backscattered light
    - energetic electrons

$I_{laser} = 2 - 8 \times 10^{14}$ W/cm^2
$\lambda_{laser} = 351$nm,
$T_e = 2.75$ keV,
$T_i = 1$ keV, Z=1,
t_{max} up to 20 ps
Length = 1.5 mm
Density profiles from NIF hydro simulations



14 million particles
~100 CPU hours per run
~1 hr on modest size local cluster

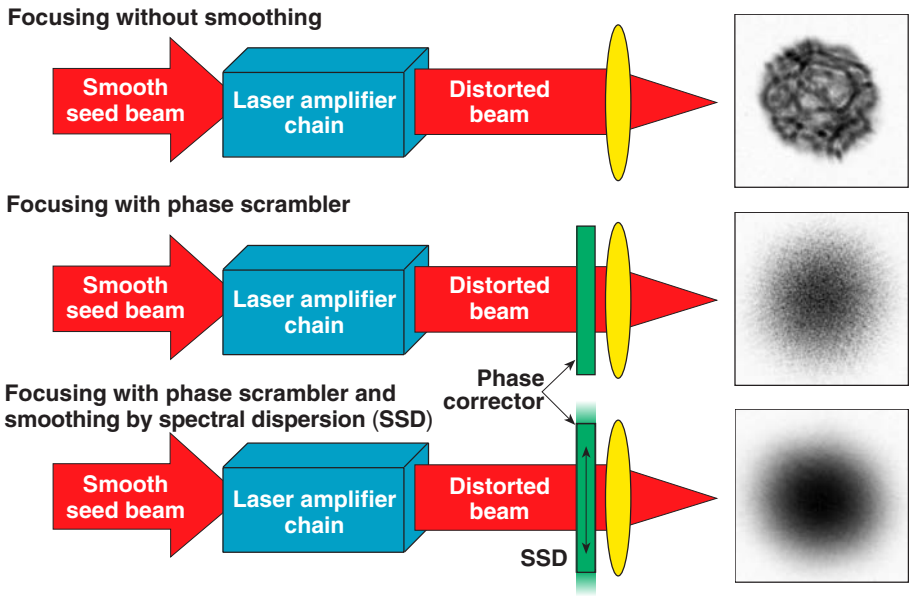


We have simulated stimulated Raman scattering in multi-speckle scenarios (in 2D)

- Although the SRS problem is 1D (i.e., the instability grows along the direction of laser propagation). The SRS problem in IFE is not strictly 1D -- each "beam" (right) is made up of 4 lasers, called a NIF "quad," and each laser is not a plane wave but contains "speckles," each one a few microns in diameter. These hotspots are problematic because you can have situations where according to linear theory, the "averaged" laser is LPI unstable only inside these "hotspots" (and the hotspots can move in time by adding colors near the carrier frequency). And the LPI's in these hotspots can trigger activities elsewhere. The multi-speckle problem are inherently 2D and even 3D.

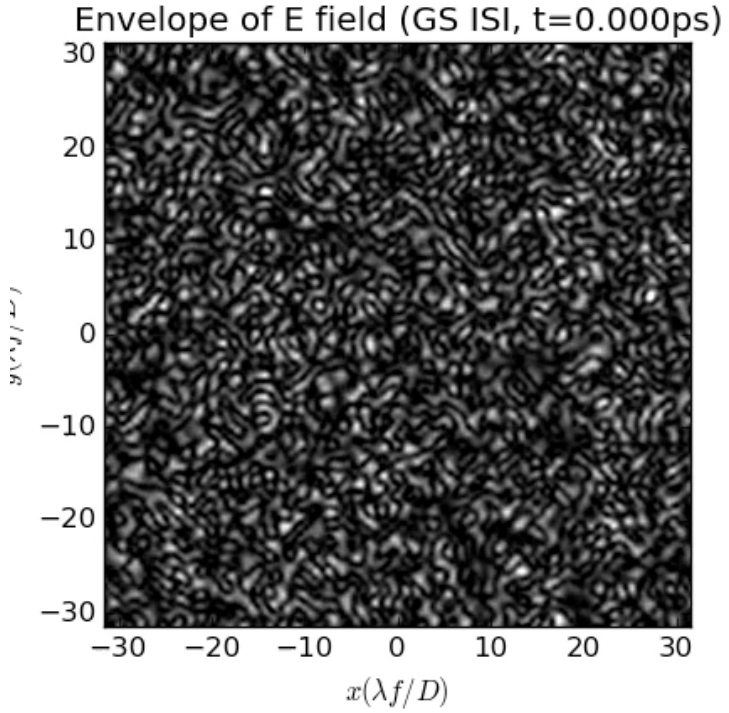
- We have been using OSIRIS to look at SRS in multi-speckle scenarios. In our simulations we observed the excitation of SRS in under-threshold speckles via:
  - "seeding" from backscatter light from neighboring speckles
  - "seeding" from plasma wave seeds from a neighboring speckle.
  - "inflation" where hot electrons from a neighboring speckle flatten the distribution function and reduce plasma wave damping.

- Recently experiments have shown that external magnetic fields can reduce LPI activities. This is another area of active research in our group.

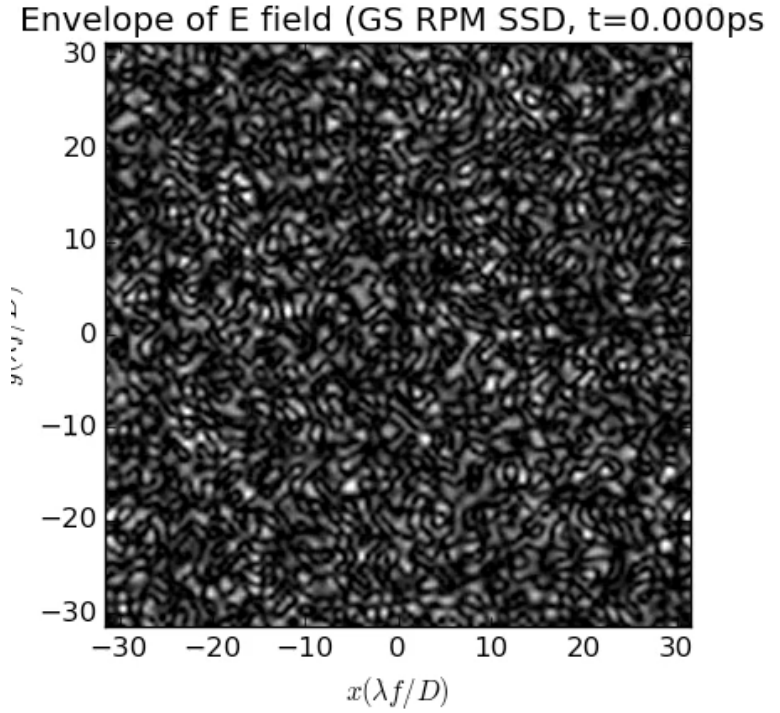


In the past year, we have added realistic beam effects into OSIRIS

ISI
(Induced Spatial Incoherence)



SSD
(Smoothing by Spatial Dispersion)



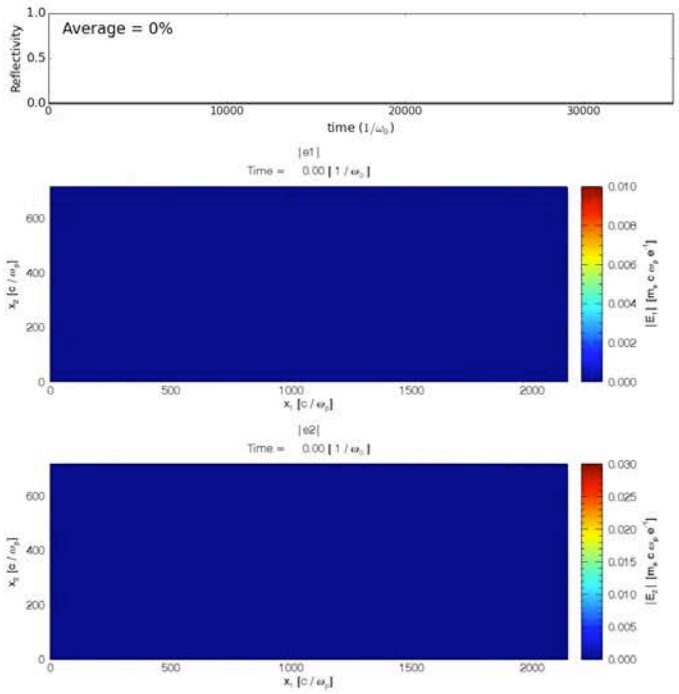
& STUD
(Spike Train of Uneven Duration)

LPI Simulation Results — Temporal bandwidth reduces LPI

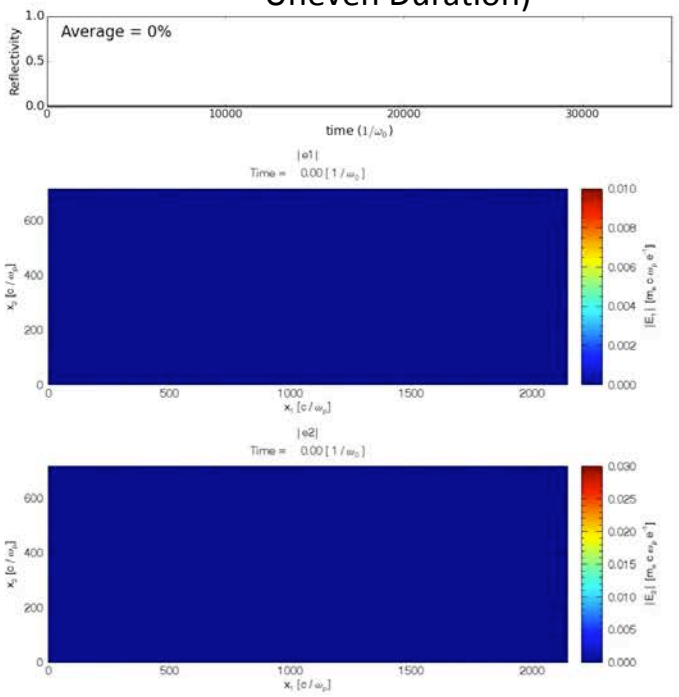
Temporal bandwidth can suppress SRS growth

- Small simulations (90k core-hours each) to identify interesting parameters before starting full simulations (<1 million core-hours each)
  - 15 speckles across and ~120 microns long.
  - ~100 million grids and ~10 billion particles each.
- Incorporating polarization smoothing can further reduce SRS reflectivity

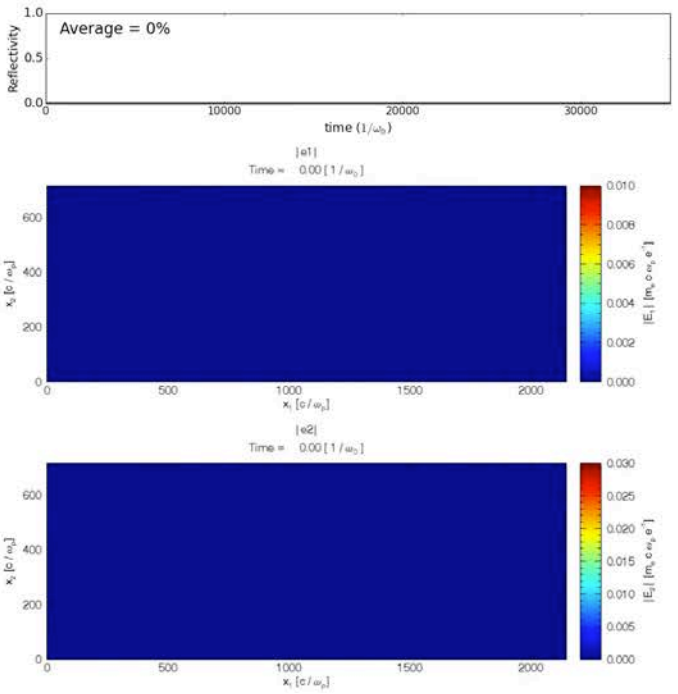
RPP



STUD (Spike Train of Uneven Duration)



ISI



PIC simulations of 3D LPI's is still a challenge, and requires exa-scale supercomputers, this will require code developments in both new numerical methods and new codes for new hardwares

	2D multi-speckle along NIF beam path	3D, 1 speckles	3D, multi-speckle along NIF beam path
Speckle scale	50 x 8	1 x 1 x 1	10 x 10 x 5
Size (microns)	150 x 1500	9 x 9 x 120	28 x 28 x 900
Grids	9,000 x 134,000	500 x 500 x 11,000	1,700 x 1,700 x 80,000
Particles	300 billion	300 billion	22 trillion
Steps	470,000 (15 ps)	540,000 (5 ps)	540,000 (15 ps)
Memory Usage*	7 TB	6 TB	1.6 PB
CPU-Hours	8 million	13 million	1 billion (2 months on the full Blue Waters

Designing New Particle-in-Cell (PIC) Algorithms on GPU's

On the GPU (and multi-cores), we apply a local domain decomposition scheme based on the concept of tiles.

Particles ordered by tiles, varying from 2 x 2 to 16 x 16 grid points (typical tile size is 16 x 16 in 2D and 8 x 8 x 8 in 3D)

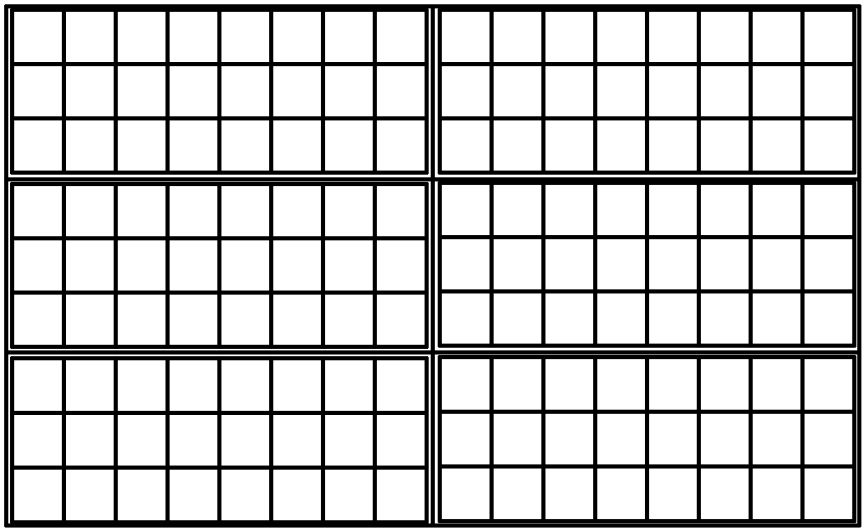
On Fermi M2090:

• On each GPU, the problem is partitioned into many tiles, and the code associate a **thread block** with each tile and particles located in that tile

We created a new data structure for particles, partitioned among threads blocks (i.e., particles are sorted according to its tile id, and there is a local domain decomposition within the GPU), within the tile the grid and the particle data are aligned and the loops can be easily parallelized.

We created a new data structure for particles, partitioned among threads blocks:

```
dimension part(npmax,idimp,num_blocks)
```



UCLA

Porting PIC codes to GPU's (continued)

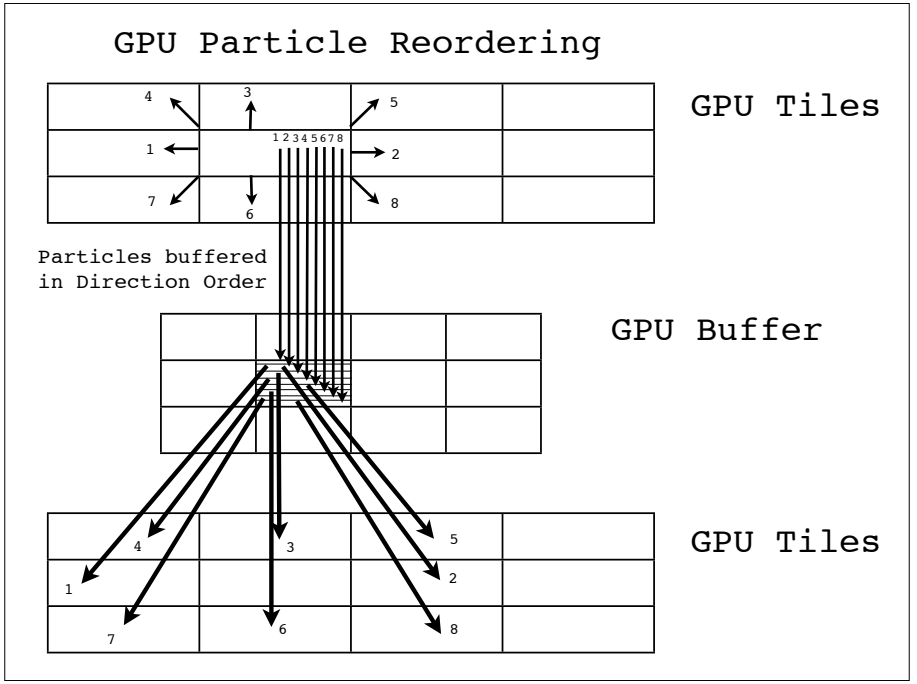
Designing New Particle-in-Cell (PIC) Algorithms:

Maintaining Particle Order

Three steps:
1. Particle Push creates a list of particles which are leaving a tile
2. Using list, each thread places outgoing particles into an ordered buffer it controls
3. Using lists, each tile copies incoming particles from buffers into particle array

A "particle manager" is needed to maintain the data alignment. This is done every timestep.
- Less than a full sort, **low overhead** if particles already in correct tile
- **Essentially message-passing**, except buffer contains multiple destinations

In the end, the particle array belonging to a tile has no gaps
- Particles are moved to any existing holes created by departing particles
- If holes still remain, they are filled with particles from the end of the array



UCLA

Evaluating New Particle-in-Cell (PIC) Algorithms on GPU: **Electromagnetic Case**
2-1/2D EM Benchmark with 2048x2048 grid, 150,994,944 particles, 36 particles/cell
optimal block size = 128, optimal tile size = 16x16

GPU algorithm also implemented in OpenMP

Hot Plasma results with dt = 0.04, c/vth = 10, relativistic

	CPU:Intel i7	GPU:Fermi M2090	OpenMP(12 CPUs)
Push	66.5 ns.	0.426 ns.	5.645 ns.
Deposit	36.7 ns.	0.918 ns.	3.362 ns.
Reorder	0.4 ns.	0.698 ns.	0.056 ns.
Total Particle	103.6 ns.	2.042 ns.	**9.062 ns (11.4x speedup).**

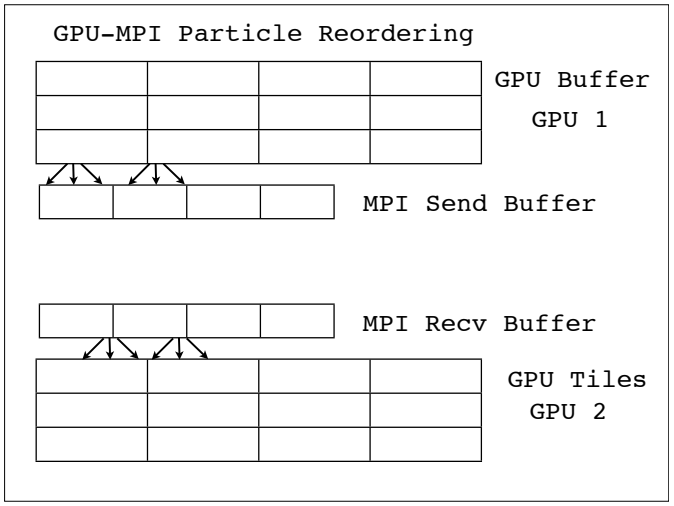
The time reported is per particle/time step.
The total particle speedup on the Fermi M2090 was 51x compared to 1 Intel i7 core.

Field solver takes an additional 10% on GPU, 11% on CPU.

OK, so how about multiple CPU/GPU's?

UCLA

Evaluating New Particle-in-Cell (PIC) Algorithms on GPU: **Electrostatic Case**
2D ES Benchmark with 2048x2048 grid, 150,994,944 particles, 36 particles/cell
optimal block size = 128, optimal tile size = 16x16 (8x8x8 in 3D). Single precision. Fermi M2090 GPU



Hot Plasma results with dt = 0.1

	CPU:Intel i7	1 GPU	24 GPUs	108 GPUs
Push	22.1 ns.	0.327 ns.	13.4 ps.	3.46 ps.
Deposit	8.5 ns.	0.233 ns.	11.0 ps.	2.60 ps.
Reorder	0.4 ns.	0.442 ns.	19.7 ps.	5.21 ps.
Total Particle	31.0 ns.	1.004 ns.	49.9 ps.	13.10 ps.

The time reported is per particle/time step.
The total particle speedup on the 108 Fermi M2090s compared to 1 GPU was 77x (>70% efficient),

We feel that we can improve on the current efficiency. Currently, field solver (which uses FFT) takes an additional 5% on 1 GPU, 45% on 2 GPUs, and 73% on 108 GPUs (there are 3 GPUs per node on our cluster, sharing one ethernet port). And we believe the efficiency should be higher for PIC codes with a finite-difference solver.

We are also working on a Intel Phi version and a OpenMP + MPI version (which achieved ~80% efficiency on 50,000 cores on Edison @ NERSC)!

here are available at the **UCLA PICKSC web-site**

UCLA

http://picksc.idre.ucla.edu/

UCLA Particle-in-Cell and Kinetic Simulation Software Center (PICKSC), NSF funded center whose Goal is to provide and document parallel Particle-in-Cell (PIC) and other kinetic codes.

http://picksc.idre.ucla.edu/
github: UCLA Plasma Simulation Group

Planned activities

- Provide parallel skeleton codes for various PIC codes on traditional and new hardware systems.
- Provide MPI-based production PIC codes that will run on desktop computers, mid-size clusters, and the largest parallel computers in the world.
- Provide key components for constructing new parallel production PIC codes for electrostatic, electromagnetic, and other codes.
- Provide interactive codes for teaching of important and difficult plasma physics concepts
- Facilitate benchmarking of kinetic codes by the physics community, not only for performance, but also to compare the physics approximations used
- Documentation of best and worst practices for code writing, which are often unpublished and get repeatedly rediscovered.
- Provide some services for customizing software for specific purposes (based on our existing codes)

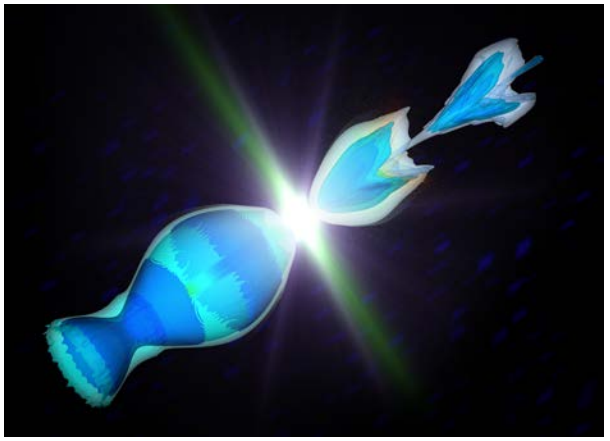
Key components and codes will be made available through standard open source licenses and as an open-source community resource, contributions from others are welcome.

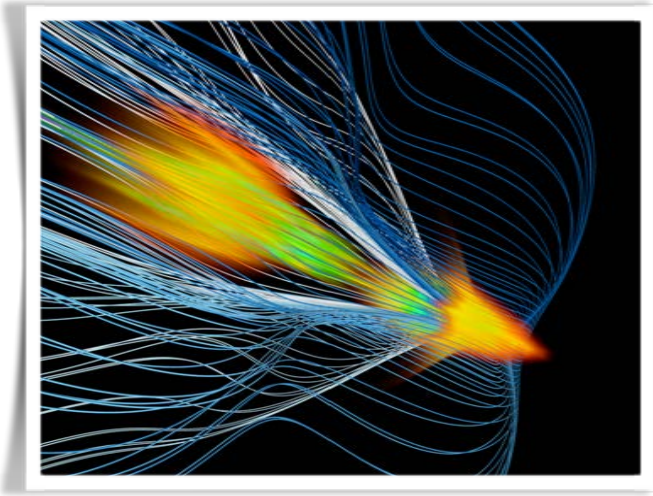
PICKSC UCLA

Summary and Outline

OUTLINE/SUMMARY

- Overview of the project
  - Particle-in-cell method
  - Our main production code — OSIRIS
- Application of OSIRIS to plasma based accelerators:
  - Production of high quality electrons for ultra-bright XFEL light sources useful for nuclear science (Xu *et al*, being prepared for *Nature Physics* (2017)).
- Higher (2 & 3) dimension simulations of LPI's relevant to laser fusion
  - Adding realisms in 2D LPI simulations relevant to laser fusion.
  - Controlling LPI's by temporal incoherence.
  - Estimates of large scale 3D LPI simulations (& justify the need for new new algorithms on exascale supercomputers)
- Code Developments to reduce simulation time and to move toward exa-scale.
  - Code developments toward exa-scale (multi-GPU's and OpenMP/MPI PIC codes)





UCLA