Computational Requirements and Milestones

Andreas S. Kronfeld (Fermilab)
Spokesperson, USQCD Collaboration

LQCD extension III Review
Cambria Rockville Hotel | July 9–10, 2019

Charge

- The critical issues to be examined in the July 9-10 review include:
  - What is the scientific case for continuing simulations of Quantum Chromodynamics (QCD) in high energy physics past 2019? Are the goals of the proposed research program aligned with the experimental and theoretical physics goals of HEP for the period 2020-2024?
  - What is the impact and interplay of lattice QCD simulations on the experimental and theoretical programs of HEP? Will the value of our experimental and theoretical programs be measurably enhanced by such simulations? Give specific examples where LQCD calculations impact the experimental program and add value to its experimental results.

Science talks & second half of this one

- The critical issues to be examined in the July 9-10 review include (cont'd):
    - Why is an extended project needed if the Office of Advanced Scientific Computing Research is providing the lattice community access to Leadership Class machines? In particular, is mid-scale hardware, such as CPU or GPU Institutional Clusters, essential and cost effective in such an environment? What is the optimal mix of machines, Leadership Class and mid-scale clusters, given realistic budget scenarios?

first half of this talk

- What are the plans at Fermilab and Brookhaven for LQCD Institutional Cluster computing? How are these plans incorporated into your proposal for the LQCD research program in 2020-2024?

next talk

Computing Requirements

Computing Landscape

- Some definitions (from wikipedia entry on supercomputers):
  - capability computing uses the maximum computing power to solve a single large problem in the shortest amount of time—
    - for example, the ANL and ORNL leadership-class facilities (LCFs);
  - capacity computing uses efficient cost-effective computing power to solve a few somewhat large problems or many small problems—
    - for example, the LQCD clusters at BNL, Fermilab, and Jefferson Lab;
  - systems supporting many users for routine everyday tasks may have a lot of capacity but are not typically considered supercomputers—
    - for example, web or email servers.

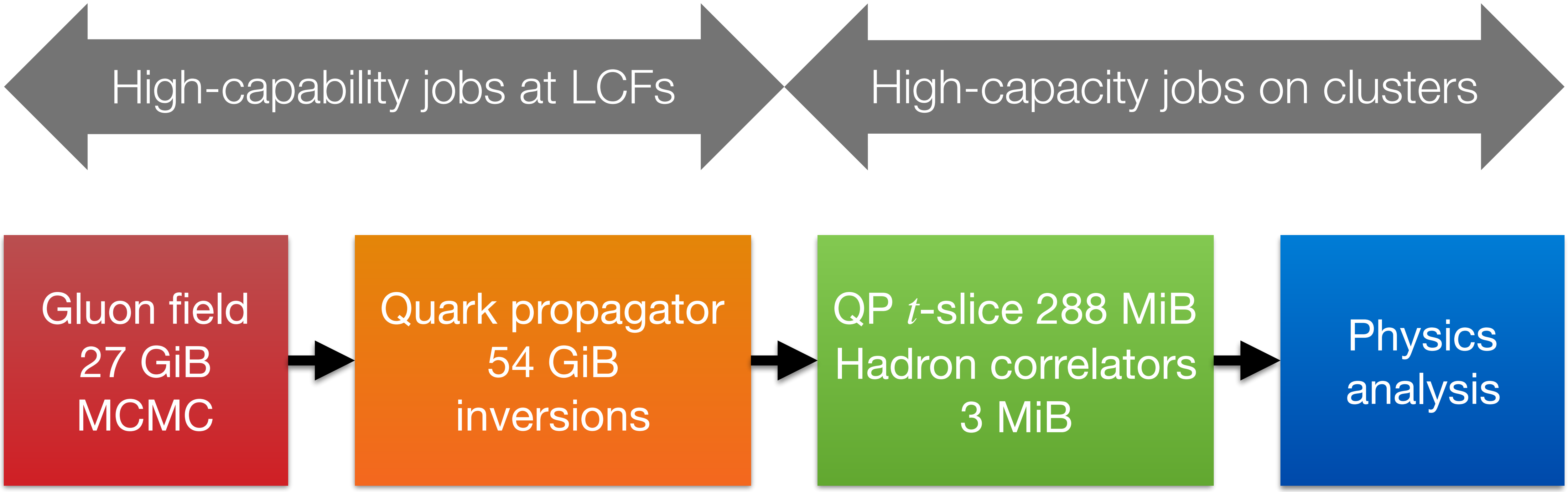
Computing Landscape

- Some definitions (from wikipedia entry on supercomputers):
    - capability computing uses the maximum computing power to solve a single large problem in the shortest amount of time—
        - for example, the ANL and ORNL leadership-class facilities (LCFs);

How do these types of computing fit into lattice-QCD workflow?

        - for example, the LQCD clusters at BNL, Fermilab, and Jefferson Lab;
    - systems supporting many users for routine everyday tasks may have a lot of capacity but are not typically considered supercomputers—
        - for example, web or email servers.

Example Lattice-QCD Workflow: Muon g–2 HVP



~10^3 gauge fields ×
dozen ensembles
= ~10^4 total

~10^4 gauge fields ×
(16–48) time sources ×
2 smearings
= ~10^6 propagator files

~10^6 propagators ×
3 directions ×
2 sources × 4 moments
= ~10^7 small files

~10^7 data points
w/ $a \neq 0$, $L < \infty$
⇒ EFT fit

from Ruth Van de Water

When needed, move files between facilities with ESNet/Globus

Example Lattice-QCD Workflow: Muon g–2 HVP

High-capability jobs at LCFs ⟷ High-capacity jobs on clusters

Gluon field
27 GiB
MCMC
→
Quark propagator
54 GiB
inversions
→
QP t-slice 288 MiB
Hadron correlators
3 MiB
→
Physics
analysis

~10^3 gauge fields ×
dozen ensembles
= ~10^4 total

~10^4 gauge fields ×
(16–48) time sources ×
2 smearings
= ~10^6 propagator files

~10^6 propagators ×
3 directions ×
2 sources × 4 moments
= ~10^7 small files

~10^7 data points
w/ $a \neq 0, L < \infty$
⇒ EFT fit
↓
a_μ^{HVP}

from Ruth Van de Water

When needed, move files between facilities with ESNet/Globus

Hardware ⊕ Policy

- In hardware, the distinction capability vs capacity can be blurry:
  - low-latency clusters have enough capability for small and medium-sized lattices; *latency = time to start communications*
  - machines like Mira (IBM® BG/Q), Theta (Intel® Knights Landing), Cori (Intel® Haswell & KNL), etc., could be used for high-throughput jobs albeit less cost effectively than farms or clusters (costly interconnects).
- Distinction can be sharpened via policies (in force at the LCFs):
  - favor proposals that require a large fraction of the available capability;
  - queues are configured to favor such jobs.
- Distinction blurred by tools to bundle many logical jobs into one superjob.

Hardware ⊕ Policy

- In hardware, the distinction capability vs capacity can be blurry:
  - low-latency clusters have enough capability for small and medium-sized lattices; *latency = time to start communications*
  - machines like Mira (IBM® BG/Q), Theta (Intel® Knights Landing), Cori (Intel® Haswell & KNL), etc., could be used for high-throughput jobs albeit less cost effectively than farms or clusters (☚ costly interconnects).
- Distinction can be sharpened via policies (in force at the LCFs):
  - favor proposals that require a large fraction of the available capability;
  - queues are configured to favor such jobs.
- Distinction blurred by tools to bundle many logical jobs into one superjob.

Further Useful Distinction

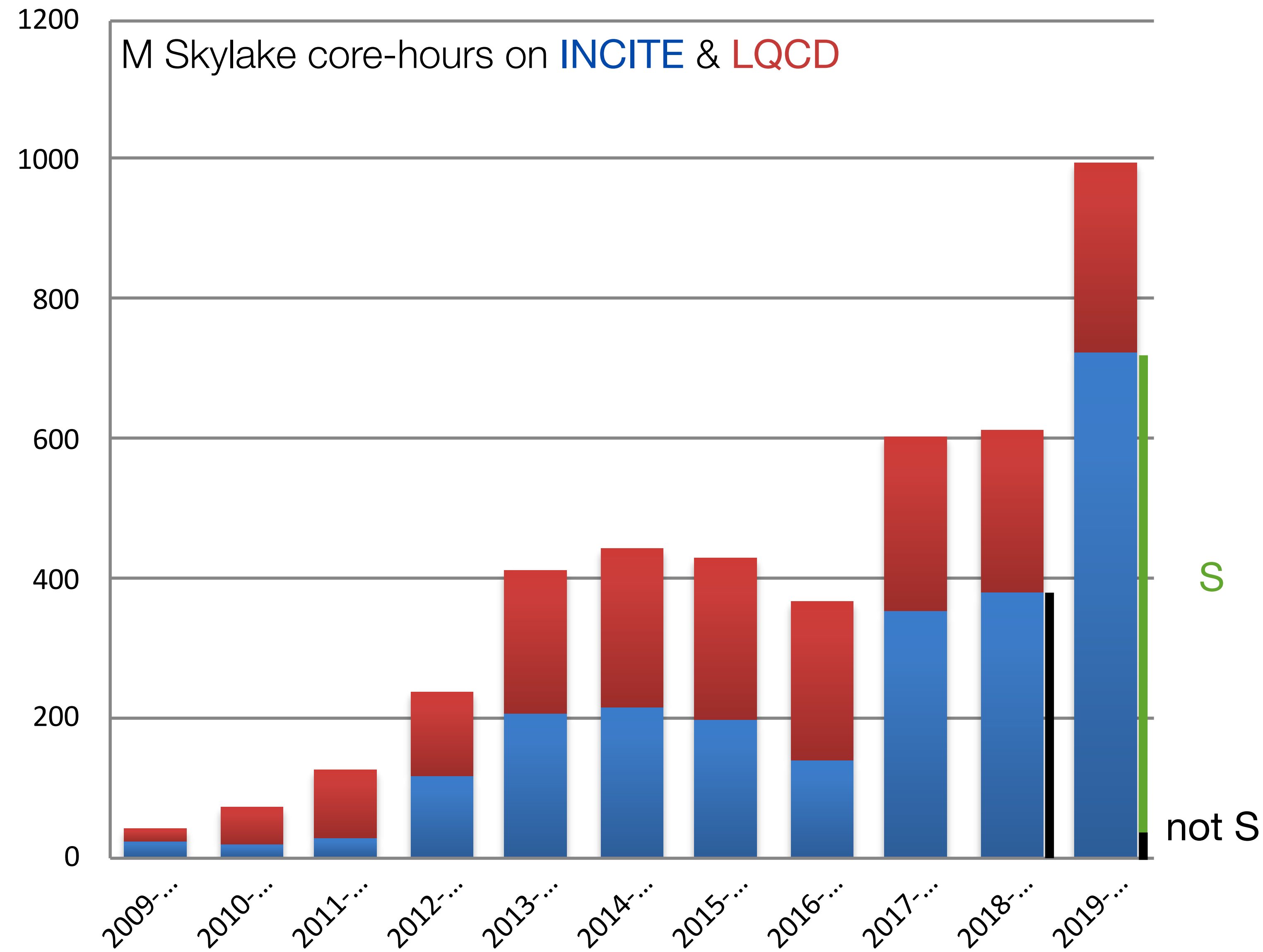
- Industrial vs. innovative.
- Industrial means the problem and its workflow are well-understood:
  - a capability job has to be industrial, otherwise not ready for LCF;
  - bundled capacity jobs will also be those that can be industrialized.
- Innovative means the problem is under development in some way:
  - new idea requiring non-trivial simulations to make progress;
  - data analysis to estimate systematic uncertainties;
  - in both cases rapid turnaround enables frequent human interaction.

Further Useful Distinction

- Industrial vs. innovative.
- Industrial means the problem and its workflow are well-understood:
  - a capability job has to be industrial, otherwise not ready for LCF;
  - bundled capacity jobs will also be those that can be industrialized.
- Innovative means the problem is under development in some way:
  - new idea requiring non-trivial simulations to make progress;
  - data analysis to estimate systematic uncertainties;
  - in both cases rapid turnaround enables frequent human interaction.

Combination of LCF and Cluster Computing

- Historical mix shown here ☛
- Over LQCD ext. II, clusters (HEP + NP) were 35–60% of the total available to USQCD.*
- Even with Summit turning on in 2018, clusters are still 25% of total.
- In 2019, USQCD received 12.8% of Summit; estimated value ~$9M.



*Additional resources are awarded to lattice QCD outside USQCD.

- With our involvement in the Exascale Computing Project, USQCD will have software to run on Summit and future LCF machines Aurora and Frontier.

machine	lab	CPU	GPU	fabric
Summit	OLCF	IBM Power9	Nvidia Volta	Infiniband
Aurora	ALCF	Intel Xeon	Intel X^e	Cray Slingshot
Frontier	OLCF	AMD EPYC	AMD Radeon	Cray Slingshot

- In collaboration with BNL and Fermilab, we will be in a position to choose the most cost-effective technology spurred by these developments.
- If competition brings prices down, especially of GPUs, cluster cost-effectiveness should exceed LCF significantly (not just moderately).

- DOE budget guidance is ~$2M/year (see Bill's talk after the break).
- Zeroing out would severely diminish productivity (*e.g.*, innovative error analysis) and creativity (innovative research towards future needs)—
    - taste in science might align more with LCFs than HEP & NP.
- Halving would potentially force us to focus on a single host lab.
- Doubling (which would bring level back to that of LQCD ext. I) is probably best achieved by ramping up: a further key to the success of the LQCD projects is the gradual but steady increase in resources.
- More modest increases could be easily absorbed in computing or with a centralized storage facility for valuable shared data.
- All computing and no students & postdocs makes science dreary (not to mention reducing productivity and innovation immensely).

Milestones and Supporting Calculations

Suggestion #3*

- The project should develop procedures to document scientific milestones uniformly over all the LQCD areas so that the project can track their annual progress quantitatively and present it more thoroughly at each review.

- *Response*: We agree with this suggestion and have developed a plan to address it. As discussed at the Review, USQCD has commissioned six whitepapers on the full range of physics topics, and a seventh on computing accomplishments and challenges. We have organized the writing in such a way to bring in enough authors to represent all our scientific goals. The whitepaper coordinators are attentive to the need to match physics relevance and computing feasibility into a set of reviewable milestones.

- At the same time, we should not formulate milestones in a way that stifles innovation or sets artificial end dates. An example of the latter is precision: just as with precision experiments, a certain target on a five-year time scale does not necessarily render irrelevant a more precise result on a ten-year time scale.

**from 2018 review of LQCD ext. II, NPPLC, USQCD.*

Considerations

- When building a detector (for example), milestones are of the type
  - design trackers with a specified momentum resolution;
  - downselect to a single design based on cost, ...;
  - build the chosen design by a specified date.
  - Collect a specified amount of integrated luminosity.
- In the dedicated-cluster model, some analogous milestones were
  - design a supercomputer with a specified performance;
  - choose hardware that most cost-effectively reaches this performance.

- Here we are interested in certain calculations that fulfill some need:
    - analogous to stating an experiment will measure something to a certain precision by a certain date—
        - vulnerable to unforeseen systematics.
- Certainly appropriate in some cases.
- Not overused, because most HEP experiments measure many quantities, and this breadth is part of the *raison d'étre*.
- In formulating milestones for a lattice-QCD research program, one must acknowledge both specificity and breadth.
- The milestones in the proposal have been formulated with these circumstances firmly in mind.

Five Milestones

- *Specific*: Complete calculations of the HVP and HLbL contributions to muon g–2 with the required precision on the timescale of Fermilab E989.
- *Specific yet broad*: Continue to sharpen the search for new physics in the quark-flavor sector.
- *Broad with many specifics*: Develop a program of precision nucleon matrix elements with comprehensive error budgets.
- *Broad but focused*: Continue the exploration of gauge theories other than QCD;
- *Broad and open*: Provide a platform for innovation on these and further relevant areas.

Supporting Calculations

Category	Milestone	Target precision	Compute %-age LCF	Compute %-age cluster	Experiment(s)
⋮					

- Excerpts from Tables I and II of the proposal:
  - "Category" groups sets of calculations according to physics;
  - "Target precision" is what we hope to achieve during LQCD ext. III;
  - "Compute %-age" is the fraction of the a) HEP-oriented part of the 2020 USQCD proposal to INCITE,* b) 2019-2020 SPC-allocated clusters;†
  - "Experiment(s)" shows the experiment(s) that profit from the calculation.
- All drawn from the USQCD whitepapers.

**adds to* 100% († *doesn't*).

Muon Anomalous Magnetic Moment

Category	Milestone	Target precision	Compute %-age LCF	Compute %-age cluster	Experiment(s)
$a_\mu = (g_\mu - 2)/2$	$a_\mu^{\text{HVP, LO}}$	0.2%	10%	4%	Muon $g-2$ (E989)
	$a_\mu^{\text{HVP, HO}}$	10%	with $a_\mu^{\text{HVP, LO}}$		Muon $g-2$ (E989)
	a_μ^{HLbL}	10%	–	2.5%	Muon $g-2$ (E989)
⋮					

- E989 target uncertainty on $10^{10} a_\mu$ is 1.6, which requires uncertainties of

HVP, LO	0.2%	700
HVP, NLO	10%	10
HLbL	10%	10

  on SM quantities. Latter two are feasible in single calculation(s).

- HVP, LO will require combining two USQCD and a few non-US results.

Quark Flavor Physics

Category	Milestone	Target precision	Compute %-age LCF	Compute %-age cluster	Experiment(s)
⋮					
CKM B physics	$f^{B\to D^{(*)}}(q^2)$	1%	10%	2%	Belle II
	$f^{B\to\pi}(q^2)$	2%	10%	5%	Belle II
	$f^{\Lambda_b\to p/\Lambda_c}(q^2)$	2%	–	–	LHCb
FCNC B physics	$f^{B\to K}(q^2)$	2%	with $f^{B\to\pi}(q^2)$		Belle II, LHCb
	$f^{B\to K^*}(q^2)$	10%*	–	–	Belle II, LHCb
	$f^{\Lambda_b\to\Lambda}(q^2)$	2%	–	–	LHCb
	$\Delta M_{B_{(s)}}$	5%*	8%	–	Belle II, LHCb, BaBar, CDF, D0
D physics	$f^{D\to\pi,K}(q^2)$	1%	with $f^{B\to\pi}(q^2)$		Belle II, BES III
K physics	$f^{K\to\pi}(0)$	0.1%	5%	–	First-row CKM unitarity
	ΔM_K	20%*	12%	2.5%	KTeV, NA48
	ε'/ε	15%	–	2.5%	KTeV, NA48
	$K\to\pi\nu\bar{\nu}$	3%	5%	1%	NA62, K0T0
⋮					

* target precision falls short of the experimental uncertainty.

- Numerous quantities, discussed on the next two slides.

- Experiments have observed several "anomalies", i.e., measurements that disagree with the Standard Model (SM):
    - a major focus of Belle II and LHCb.
- In many cases, semileptonic form factors enter the SM prediction:
    - relies on results from lattice QCD;
    - plan to use approach demonstrated for leptonic decay constants (now below the percent level, exceeding experiments' needs).
- The same form factors are used to determine CKM elements (Ruth's talk):
    - needed throughout flavor physics to know SM prediction.
- Basic QCD parameters:
    - improvements in α_s underway, in quark masses unnecessary.

- Improved results for neutral-meson mixing are needed.
- Mass differences:
  - ΔM_B is short-distance only: in mild tension with measurements;
  - ΔM_K requires novel calculations of long-distance effects, being developed further;
  - ΔM_D is long-distance dominated and extremely challenging.
- Direct *CP* violation in the kaon system:
  - first complete calculation a highlight of LQCD ext. II;
  - matching KTeV/NA48 precision during LQCD ext. III.
- With similar methods, long-distance part of the very rare decay $K^+ \to \pi^+ \nu \bar{\nu}$ (studied at BNL E949 and CERN NA62) will be computed.

Nucleon Matrix Elements

Category	Milestone	Target precision	Compute LCF	%-age cluster	Experiment(s)
⋮					
Nucleon matrix elements	Nucleon g_A^{u-d}	1%*	with $F_A(q^2)$		Neutron lifetime puzzle
	Nucleon g_T^{u-d}	1%	NP	1.5%	UCNB, Nab
	Nucleon g_S^{u-d}	3%	NP	1.5%	UCNB, Nab
	$\sigma_{\pi N}$, σ_s	5%	NP	2%	Mu2e, LZ, CDMS
	Nucleon r_E, r_A	5%	with $F_A(q^2)$		DUNE, MicroBooNE, NOvA, T2K
	Nucleon $F_A(q^2)$	8%	NP	15%	DUNE, MicroBooNE, NOvA, T2K
	Nucleon tensor	20%	NP	3%	DUNE, MicroBooNE, NOvA, T2K
	Nucleon PDFs	12%*	NP	15%	ATLAS, CMS, DUNE
	Proton decay	10%	NP	–	DUNE, HyperK
	$nn \to pp$	50%*	NP	4%	EXO, other $0\nu\beta\beta$ experiments
	Nucleon EDM	10%*	NP	3.5%	Neutron, proton EDM experiments
	$g_{A,T,S}$, $A \leq 4$	20%*	NP	3%	All neutrino, DM, EDM, …

* target precision falls short of the experimental uncertainty.

- Numerous quantities, discussed on the next three slides.

dark-matter-nucleus cross sections

muon-nucleus cross sections

neutrino-nucleus cross sections

⋮					
Nucleon	Nucleon g_A^{u-d}	1%*	with $F_A(q^2)$		Neutron lifetime puzzle
matrix	Nucleon g_T^{u-d}	1%	NP	1.5%	UCNB, Nab
elements	Nucleon g_S^{u-d}	3%	NP	1.5%	UCNB, Nab
	$\sigma_{\pi N}$, σ_s	5%	NP	2%	Mu2e, LZ, CDMS
	Nucleon r_E, r_A	5%	with $F_A(q^2)$		DUNE, MicroBooNE, NOvA, T2K
	Nucleon $F_A(q^2)$	8%	NP	15%	DUNE, MicroBooNE, NOvA, T2K
	Nucleon tensor	20%	NP	3%	DUNE, MicroBooNE, NOvA, T2K
	Nucleon PDFs	12%*	NP	15%	ATLAS, CMS, DUNE
	Proton decay	10%	NP	–	DUNE, HyperK
	$nn \to pp$	50%*	NP	4%	EXO, other $0\nu\beta\beta$ experiments
	Nucleon EDM	10%*	NP	3.5%	Neutron, proton EDM experiments
	$g_{A,T,S}$, $A \leq 4$	20%*	NP	3%	All neutrino, DM, EDM, ...

* target precision falls short of the experimental uncertainty.

- Numerous quantities, discussed on the next three slides.

- Neutrino Physics:
  - axial charge g_A to 1% as benchmark and for neutron lifetime puzzle;
  - axial form factor up to $Q^2 \lesssim 1\ \mathrm{GeV}^2$ with all uncertainties controlled and full error budget to understand quasielastic peak—
    - start with axial radius r_A to 5% (total) uncertainty;
  - exploration of best techniques for hadron tensor for SIS at larger Q^2, hoping for 20% uncertainty.
- Parton Distribution Functions
  - highest momentum transfers in νA scattering—deep inelastic;
  - LHC searches require PDFs at high Bjorken x, especially gluon PDF—
    - 12% is a useful target, improving PDF for $x \in [0.7, 0.9]$ by ~20%.

- DM Detection, Muon-to-electron Conversion, Neutron Decay:
    - share the same matrix elements, because they all entail new physics coupling to a nucleon;
    - "sigma terms", i.e., quark content of the nucleon to a few %, despite—as flavor singlets—demanding disconnected diagrams;
    - additional spin structures—scalar, axial, and tensor—needed in case Mu2e observes a signal;
    - isovector charges to 10% is very straightforward (forecast 1–3%) and with new neutron beta decay experiments reach higher energy scales for BSM interactions than LHC.

- Violations of Fundamental Symmetries:
    - baryon-number violation—proton decay and $n\bar{n}$ oscillation—will be probed with large neutrino detectors (& perhaps dedicated ESS $n\bar{n}$);
    - permanent EDMs, of neutron, proton, and nuclei, require several matrix elements—strong CP violation and/or BSM CP violation;
    - neutrinoless double-beta decay, starting with $nn \to pp$ process, with $\pi^- \to \pi^+$ as warm-up.
- Small Nuclei:
    - The program of nucleon matrix elements nuclear many-body theory (based on chiral EFT) will be complemented by calculations of properties of small nuclei, $A \le 4$. From them, deduce EFT couplings.
    - Not part of this proposal *per se*, but mentioned for context.

Strongly Coupled Gauge Theories BSM

Category	Milestone	Target precision	Compute %-age LCF	Compute %-age cluster	Experiment(s)
Higgs[a]	$\alpha_s(m_Z)$	0.3%	with quark flavor		ATLAS, CMS, FCC, ILC
+	Light spectrum	NA	10%	2%	ATLAS, CMS
BSM	Anom. dim.	NA	5%	1.5%	ATLAS, CMS
	Composite DM	NA	3%	–	LZ, CDMS
	Susy	NA	2%	1%	ATLAS, CMS
⋮					

[a] Note also the PDFs under “Nucleon Matrix Elements” on nucleon slide.

- General features of composite Higgs: spectrum and 4-quark anomalous dimension (which must be large to generate SM fermion masses).
- Spectrum and form factors of composite DM baryons, mesons, (glueballs) to deduce identity of lightest DM particle. Thermodynamics ↔ gravity waves.
- Thermodynamics of super Yang-MIlls to test conjectures about string theory.

Innovative Projects

- Flexibility fosters innovation, especially enabling new approaches, *e.g.*,
  - muon *g*–2 HVP (years ago) and HLbL (more recently);
  - x-dependent parton distribution functions (during LQCD ext. II);
  - machine-learning possibilities during LQCD ext. III.
- Nimble bottom-up approach responds to new developments.
- Career development: postdocs and junior faculty
  - try their ideas on USQCD clusters, starting with low-pressure Type B proposals, then Type A;
  - thus established, advance to NSF, NERSC, ALCC, INCITE.

Summary

Why Clusters?

- Cluster computing for lattice QCD:
  - ensures alignment of lattice-gauge-theory calculations with P5 and DOE-HEP priorities;
  - enables careful treatment of systematic uncertainties—crucial for interpreting experiments;
  - fosters innovations in new calculations and method;
  - provides platform for early career scientists to establish their reputations.

Milestones

- The milestones are designed to
  - deliver numerous lattice-QCD calculations that are key to ongoing and future experiments;
  - allows for exploration of themes emphasized by P5—mostly for experiment but also for questions of theory;
  - provide enough detail for future review panels to judge, without holding the program hostage to unforeseen circumstances;
  - be suitably ambitious yet cautious—as in the past, a few targets will be unexpectedly easy to hit, while a few others may be missed.

Questions and Discussion