Taking the lid off the National Seismic Hazard Model technical delivery - a New Zealand research project in the age of cloud computing

Chris Chamberlain and Chris DiCaprio

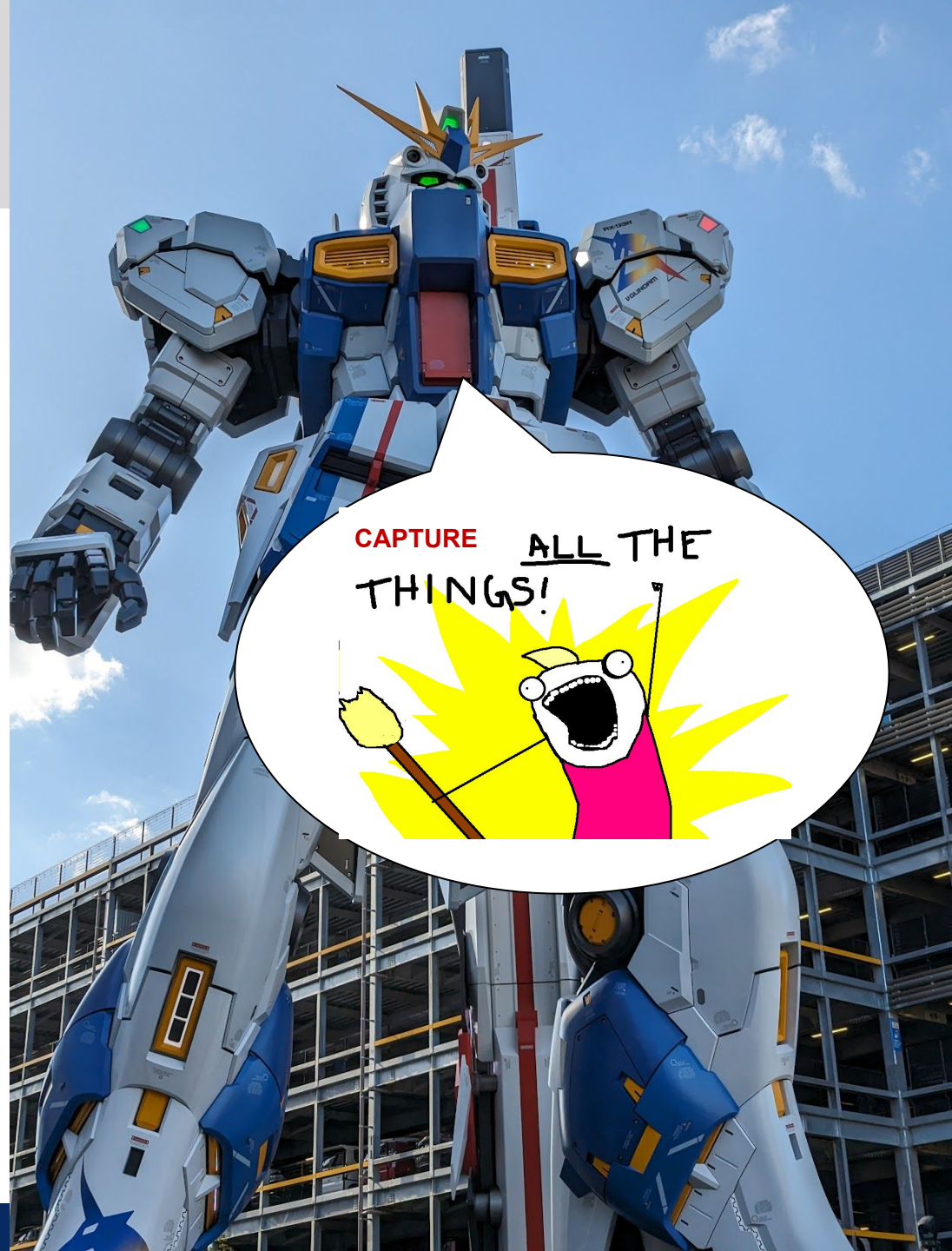
Te Tauira Matapae Pūmate Rū i Aotearoa
NSHM The New Zealand National Seismic Hazard Model
A GNS Science Led Research Programme

E mahi ana me
In collaboration with

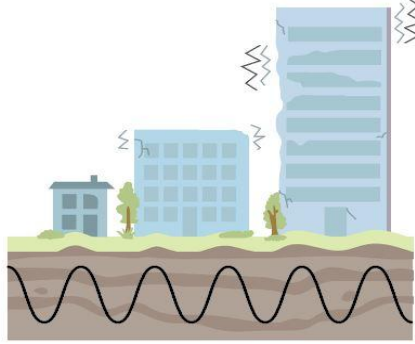
Ngā hoa tuku pūtea
Funding partners

Talk Outline..

- Introducing the NSHM
- the team and our objectives
- guiding tenets, code organism
- a story of fear and scaling
- the cloud architecture
- finding the things



National Seismic Hazard Model (NSHM)

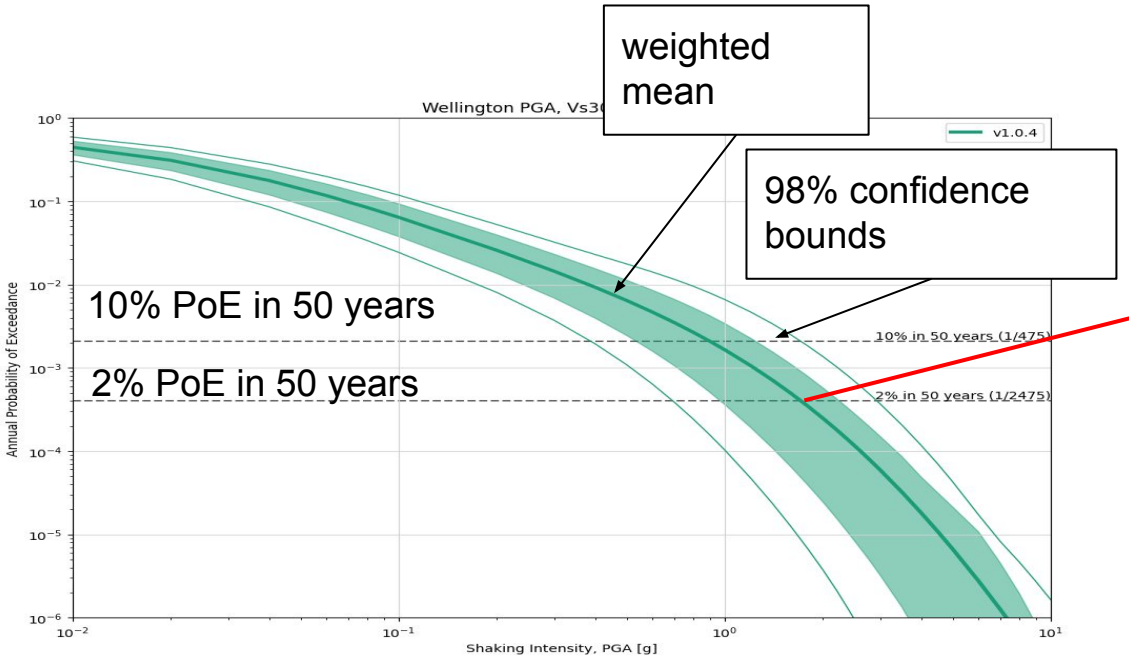


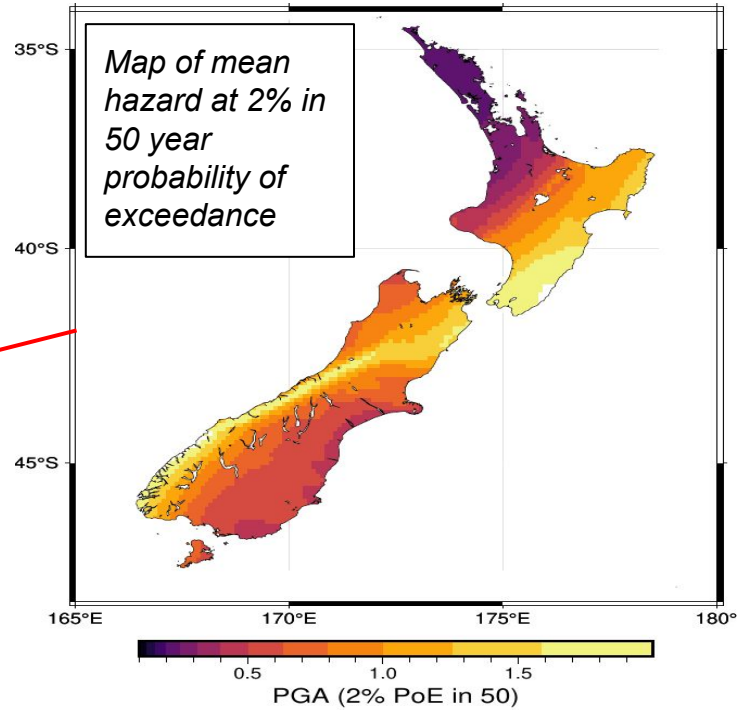
NSHM produces forecasts of shaking across all of NZ aka “seismic hazard”.

It informs a) building design and regulation, b) insurance, and c) community planning and risk management.

The new 2022 NSHM includes nearly a million models (potential futures) in order to convey the underlying uncertainty.

NSHM is comprised of thousands of models (possible futures) that are used to quantify uncertainty





The science development and review process

We aim to represent a broad range of scientific views

- with something as complex as earthquakes it is not realistic or prudent to develop a single consensus model – users need to understand the uncertainty (most want to)
- Expert selection (who is an expert?) and structured elicitation process

NSHM includes scientific understanding from around the world

- Includes a broad range of scientific views
- **More than 50 scientists from around New Zealand and around the world**
- University of Canterbury, University of Otago, University of Auckland, NIWA and others
- United States, Canada, Italy, Germany, Australia, England

NSHM Participatory peer review:

- Technical advice on the development of the NSHM has been provided by a 17-member panel of international scientists, engineers, insurance using a participatory review process.
- Scientifically detailed involvement from panel – weekly input
- Panel included key NSHM end-users
- Time consuming and challenging, but very beneficial

Assurance review:

- International review of processes: science, decision making and peer review, with positive outcomes

A tale of two Chris(es)

Chris Chamberlain - solution architect and developer (python, java, etc). IT since 1986, but a newbie to GNS and science projects.

Goals: coordinate all the software wrangling, guide the IT processes, capability, and collecting the scientific artefacts. Facilitate the public-facing NSHM.

Chris DiCaprio - Ph.D. in geophysics with experience in scientific computation (inverse methods, FEM). **Started with NO EXPERIENCE** in software architecture and cloud.

Goals: provide scientific guidance to the computational working group. Represent the team to outside groups.

Guiding IT tenets

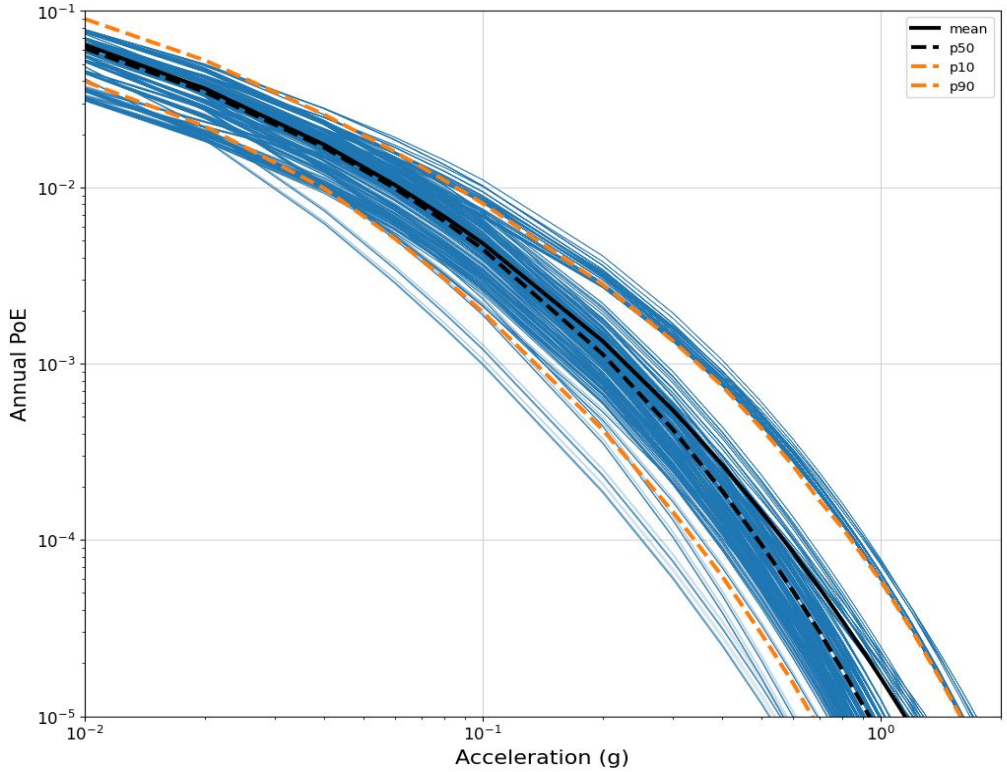
- Make our work public at https://github.com/GNS-Science
- capture all the things(TM) and make them accessible
- Minimise ‘fixed’ IT overhead, using ‘cloud’ & serverless
- Keep options open for compute intensive tasks (local, HPC, cloud)

The Hazard calculation: fear and ~~scaling~~

The new NZ NSHM is significantly bigger than any previous NSHM (many more logic tree branches). **979,776 total realizations.**

With only months remaining to keep project on schedule, it became clear that we had a serious problem computing hazard curves using openquake software

Hitting time & memory bounds … talking to GEM … are we doing it right … alternate options like change the philosophy??



Decompose the model and break calculation into 2 stages

Stage 1

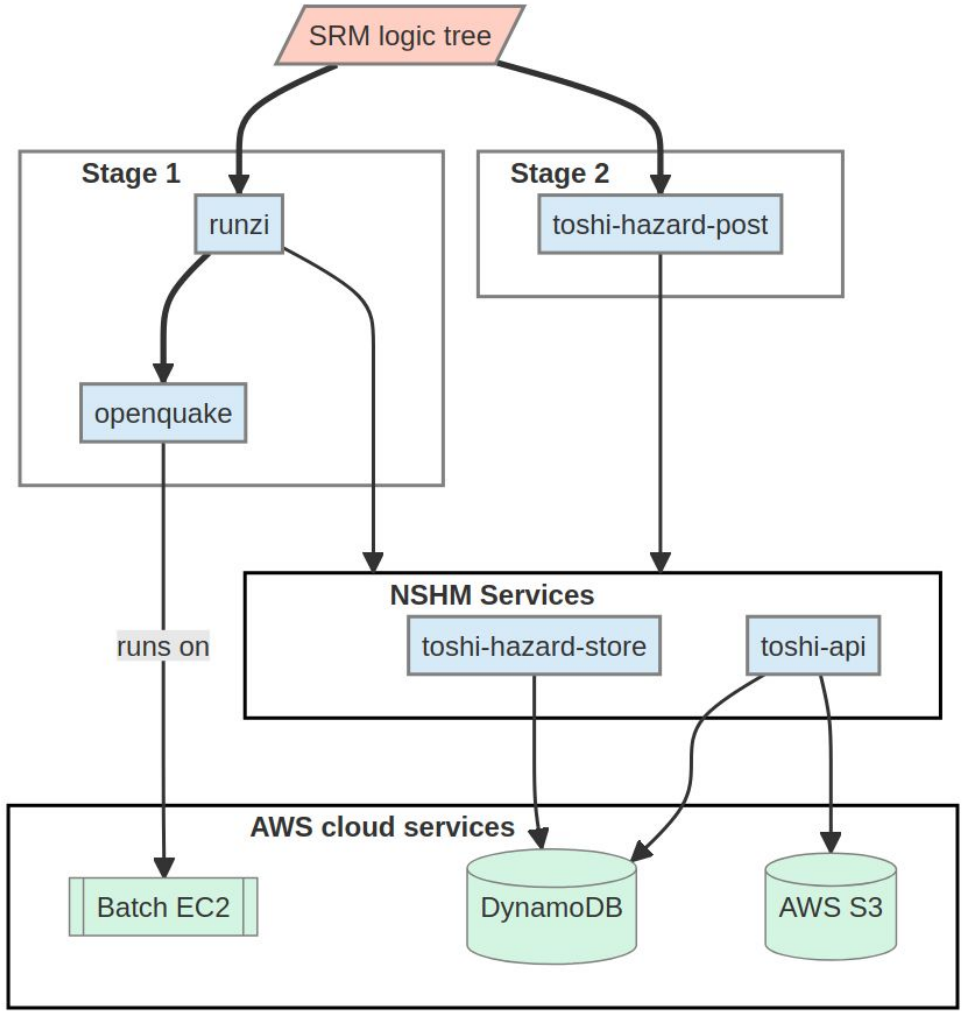
- Calculate hazard for independent components concurrently using **AWS Batch EC2**
- Store components using **AWS DynamoDB**

Stage 2

- Combine independent components to form 979,776 realizations
- Calculate aggregate statistics (e.g. weighted mean and fractiles)
- Store aggregate hazard curves in **AWS DynamoDB**

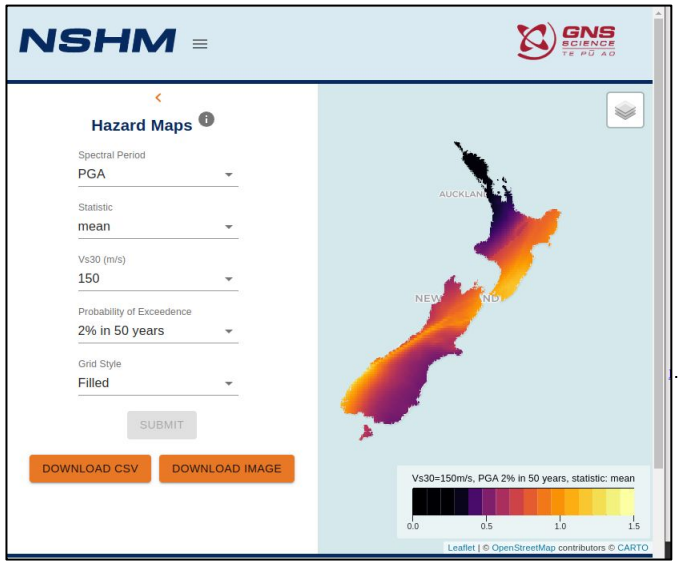
The Hazard calculation: ~~fear and~~ scaling

- Allowed us to **parallel process hazard curves on modest compute nodes ... sometimes up to 750 concurrent jobs**, each between 1 to 24 hours duration.
- Recombination and calculation of aggregate statistics is quite efficient and takes only hours running on ~40 cores
- Both stages of the process **easily parallelize and scale** across multiple dimensions of the model
- Breaking the calculation into stages and storing intermediate products provide benefits for experimentation
    - New logic tree configurations, easy sensitivity testing, sub-sampling to find minimum-viable logic trees, minimise calculation overheads, etc.

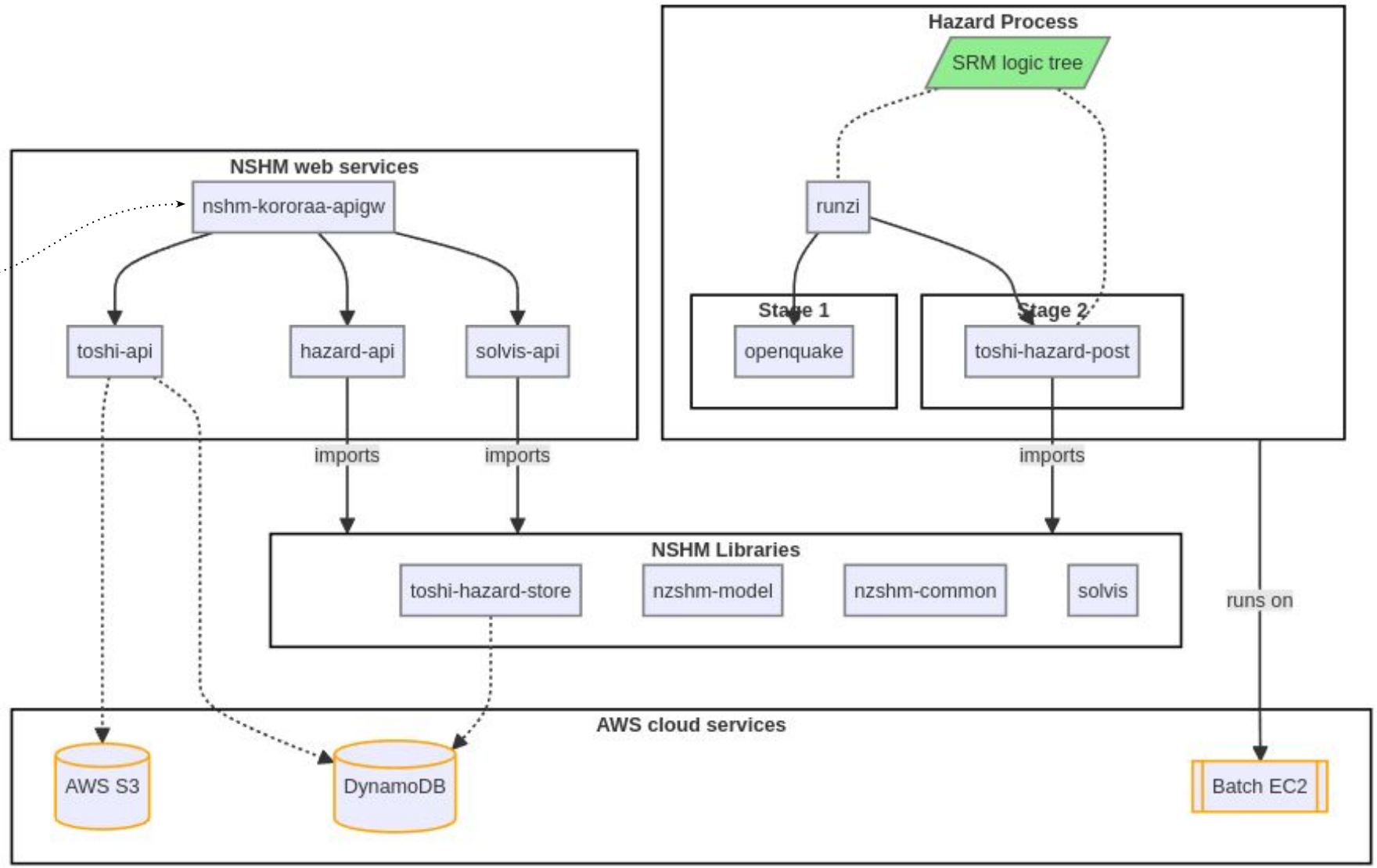


DiCaprio, et. al., Calculation of National Seismic Hazard Models with Large Logic Trees: Application to the NZ NSHM 2022. Seismological Research Letters 2023 https://doi.org/10.1785/0220230226

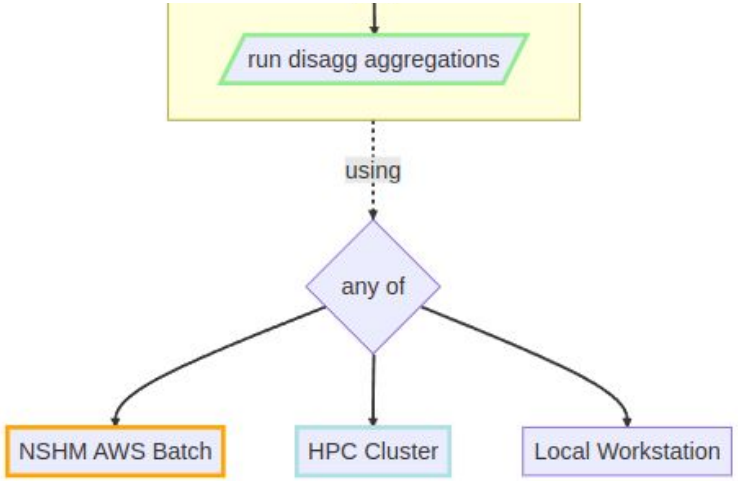
Calculating and consuming Hazard Curves

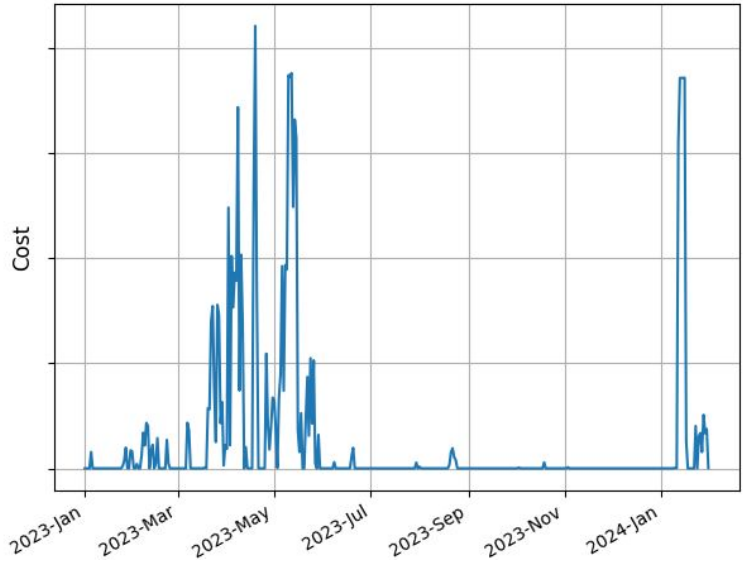


Kororā Web Application



How we use cloud services ...





Stuff we do with AWS:

- automate all experiments at scale
- store and share all the scientific data
- all the APIs are serverless (spikes)
- secure the assets
- **AWS Elasticsearch** all the API things

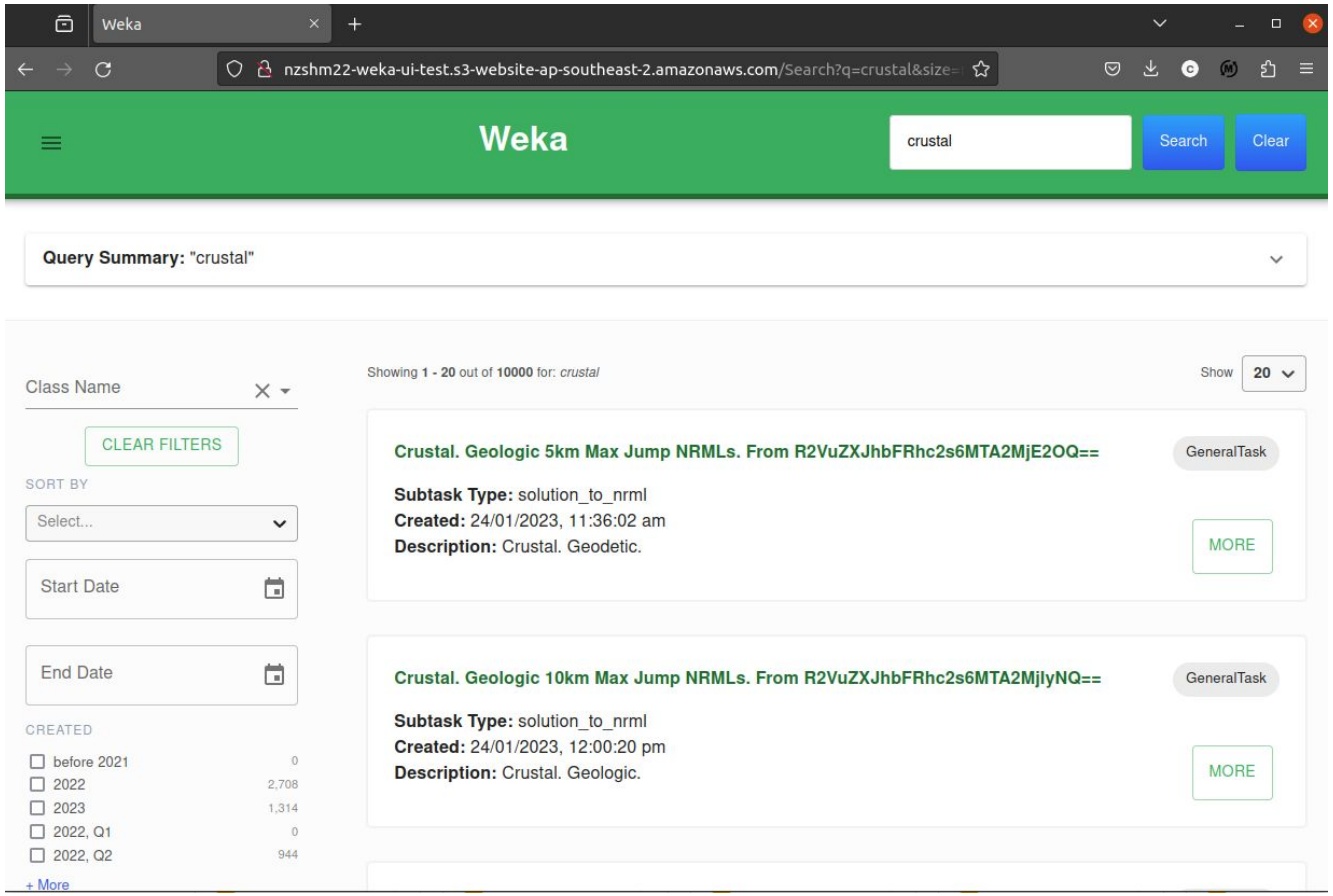
and with Github:

- automated testing and deployments
- docs and library publication

Where is everything?

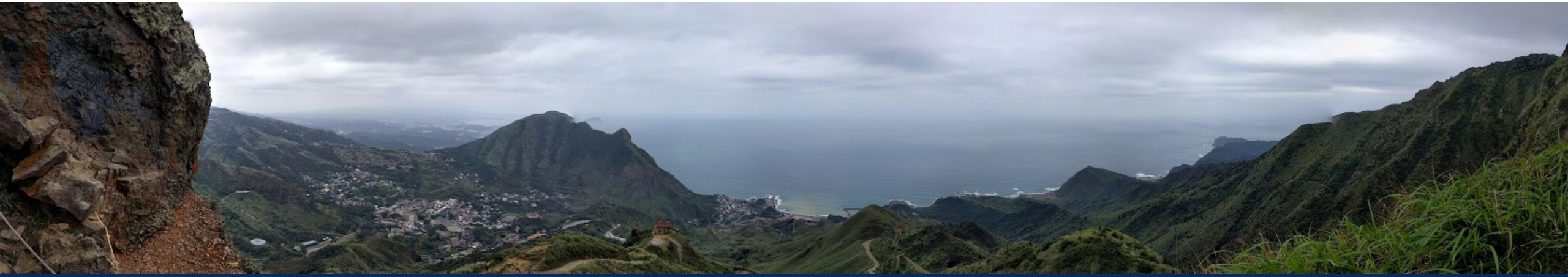
Number of

- Rupture sets generated: **234**
- Inversions run: **12,300 (322 General Tasks)**
- Hazard realizations: **119,472**
- Composite hazard curves: **128,350,656**



Next steps to July 24

- **Improving library usability for a wider user group**
  - refactoring and documentation
  - workflow options for cloud and standalone databases
- **Work with GEM** to ensure this method remains compatible with openquake developments.
- **Hazard on demand**



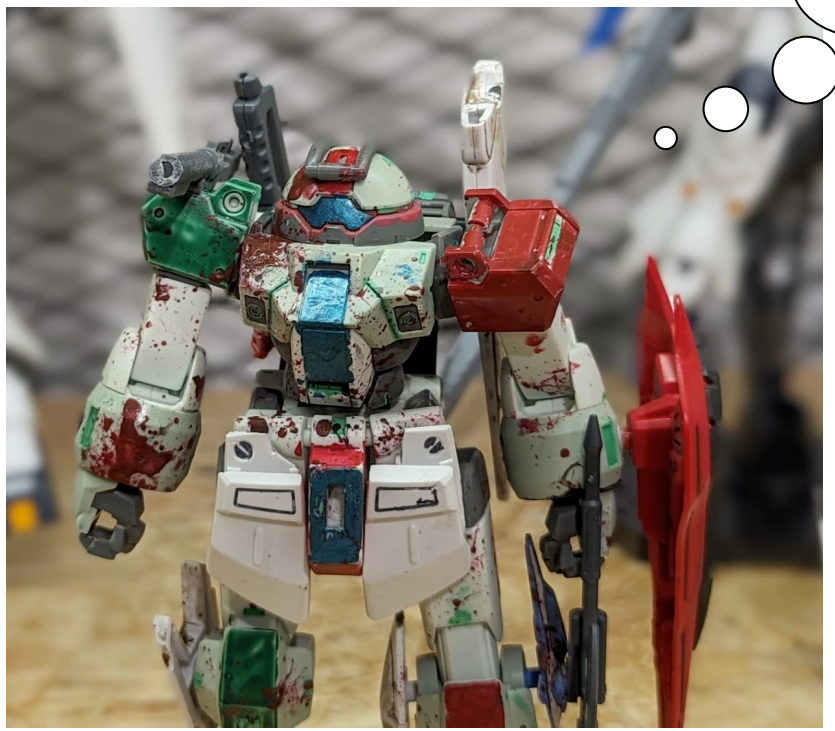
Wrap up

Cruftwork

After this slide are the discards...

Order from the chaos?

For example, finding the hazard curves...



How .... ??

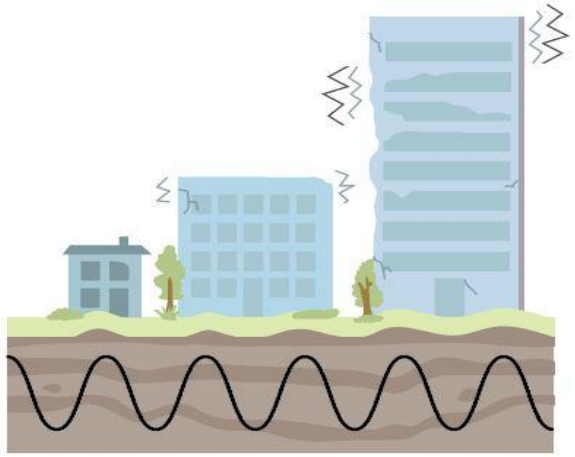
Outline

- Introduction to PSHA / NSHM - ~2 slides CDC
- NSHM Project team context
- CWG team context and challenges/unknowns ~2 sides CBC
- Doc diagrams → timeline of development and what/why went into the cloud

The NSHM produces forecasts of shaking

The National Seismic Hazard Model is a Probabilistic Seismic Hazard Analysis

- **The model provides a probabilistic forecast of earthquake shaking. This is called the seismic hazard.**
- **The model quantifies the rate of exceeding various ground motion levels given all possible earthquakes.**



The NSHM provides important input for making risk based decisions:

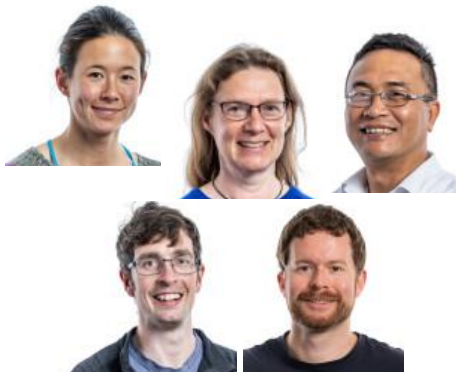
- Building regulatory system (building code)
- Risk assessments for insurance and reinsurance industry
- Community risk reduction, response, and resilience plans

Project Team and Collaboration

More than 50 scientists and specialists from around New Zealand and around the world in 3 working groups

I'm looking for an experiment we ran 2 years ago among 12,000

SRM Working Group



What if we tweaked this single parameter 96 ways?

How do we calculate this large model?

Computational Working Group

How do we make results easy to find and share?

I'd like to see a hazard curve for this specific set of inputs

I need 12 new gmm classes in the hazard calculation code

GMCM Working Group

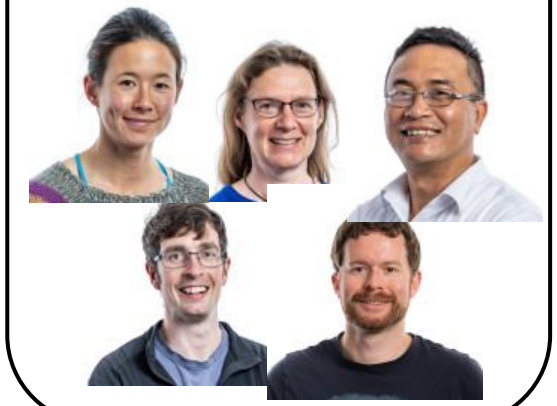
Project Team and Collaboration

More than 50 scientists and specialists from around New Zealand and around the world in 3 working groups

I'm looking for an experiment we ran 2 years ago among 12,000

What if we tweaked this single parameter 100 ways?

SRM Working Group



How can we facilitate experimentation?

How do we calculate this large model?

Computational Working Group

How do we make results easy to find and share?

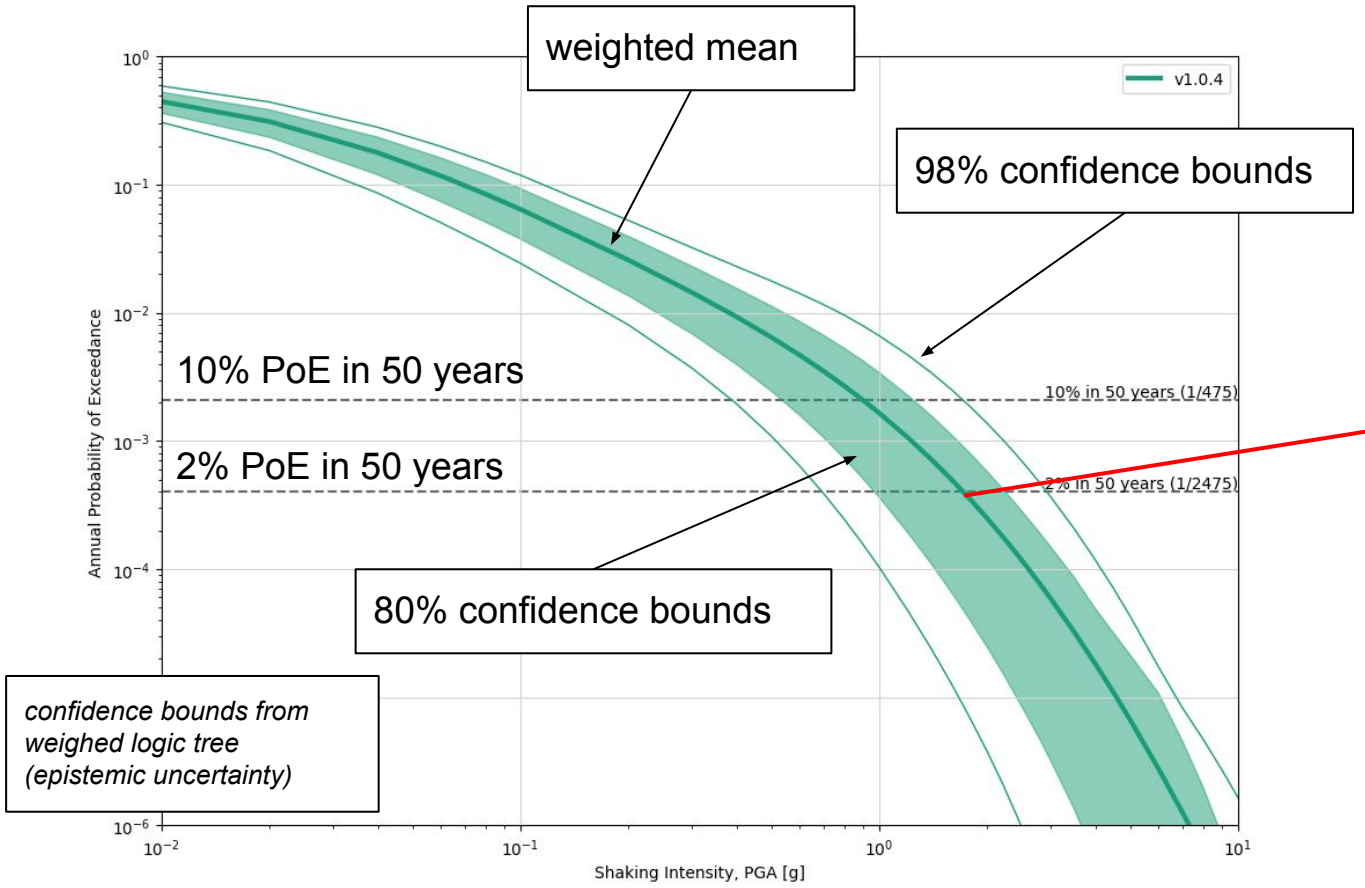
How am I going to track all of this?

I'd like to see a hazard curve for this specific set of inputs

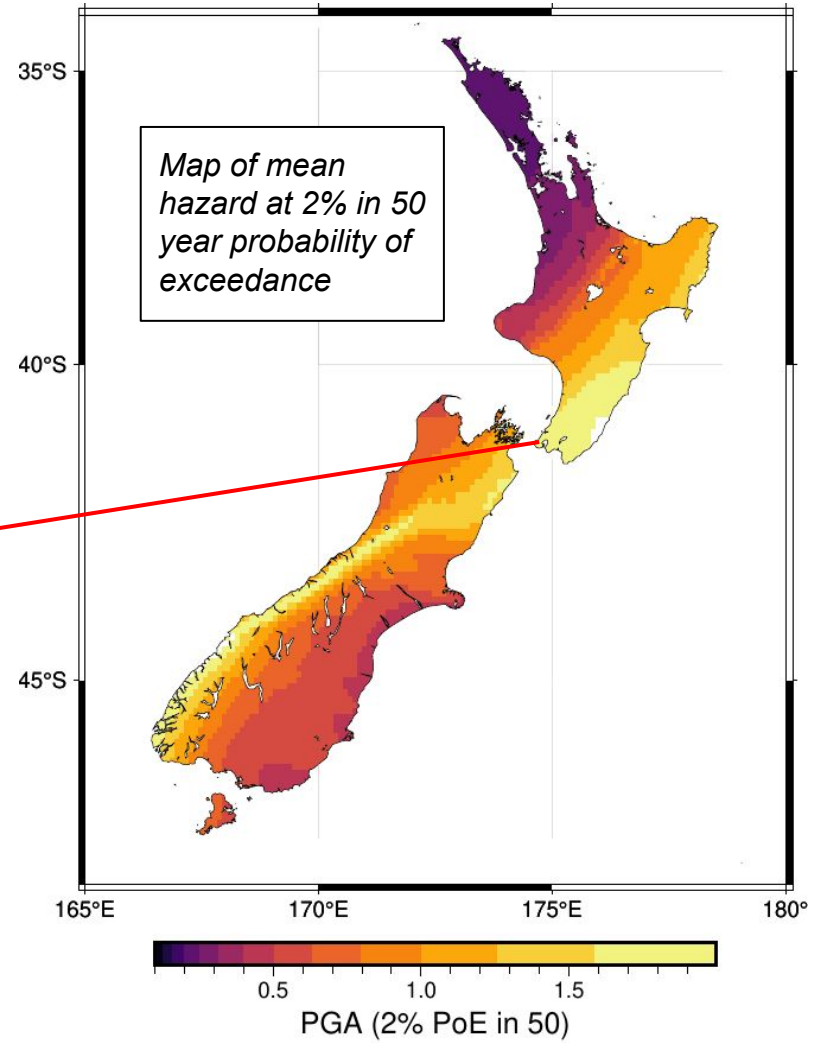
I need 12 new gmm classes in the hazard calculation code

GMCM Working Group

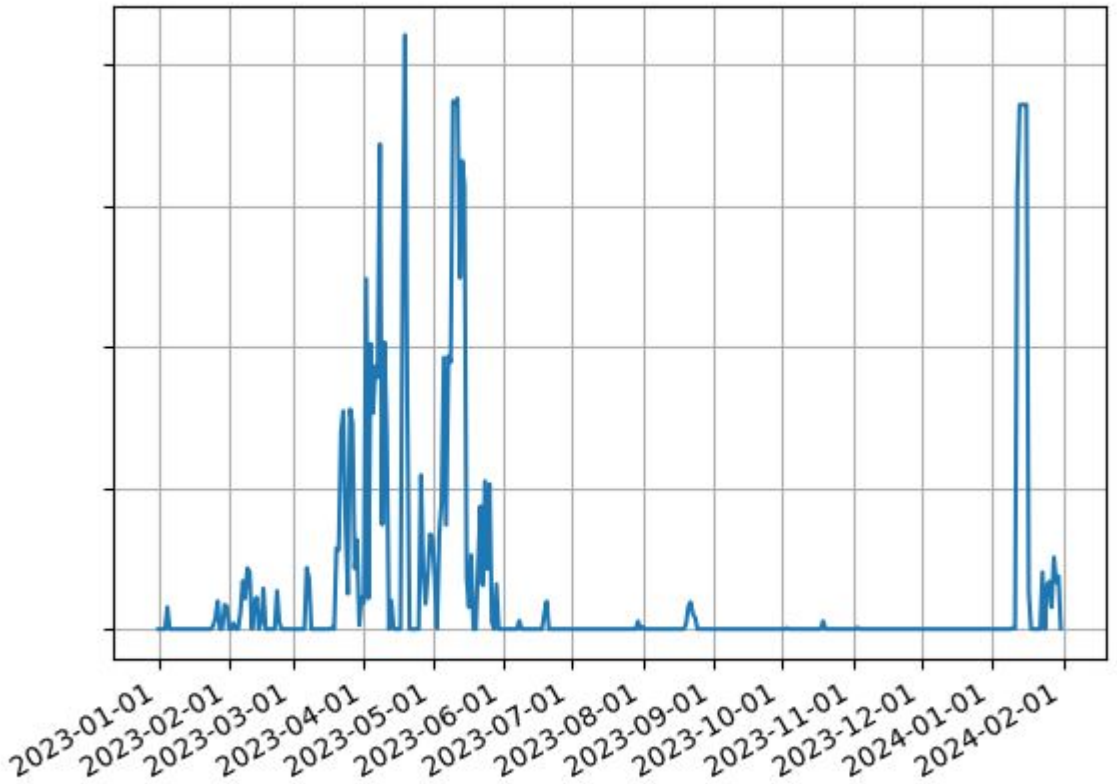
New Zealand NSHM 2022 outputs



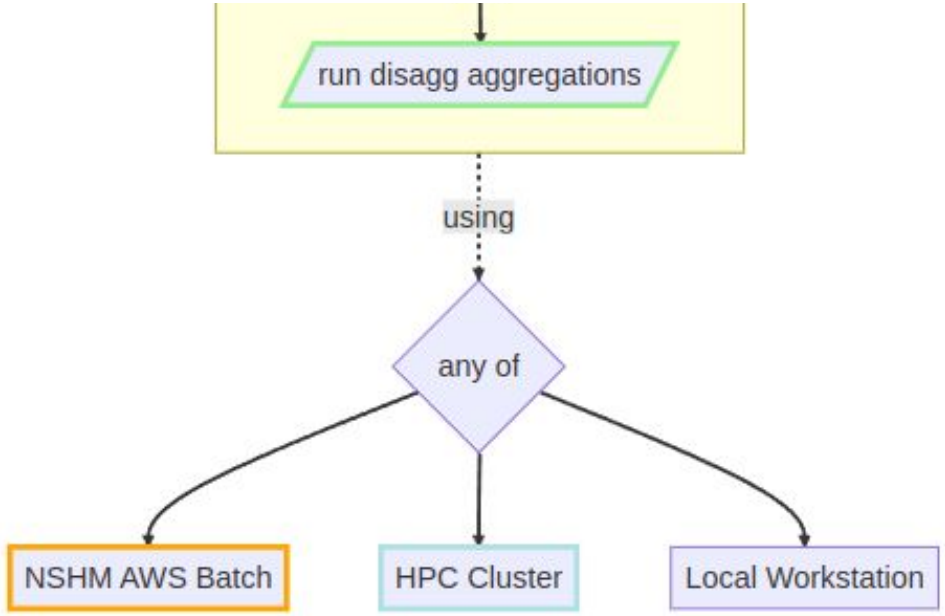
confidence bounds from weighed logic tree (epistemic uncertainty)



“Spikey” Compute Usage

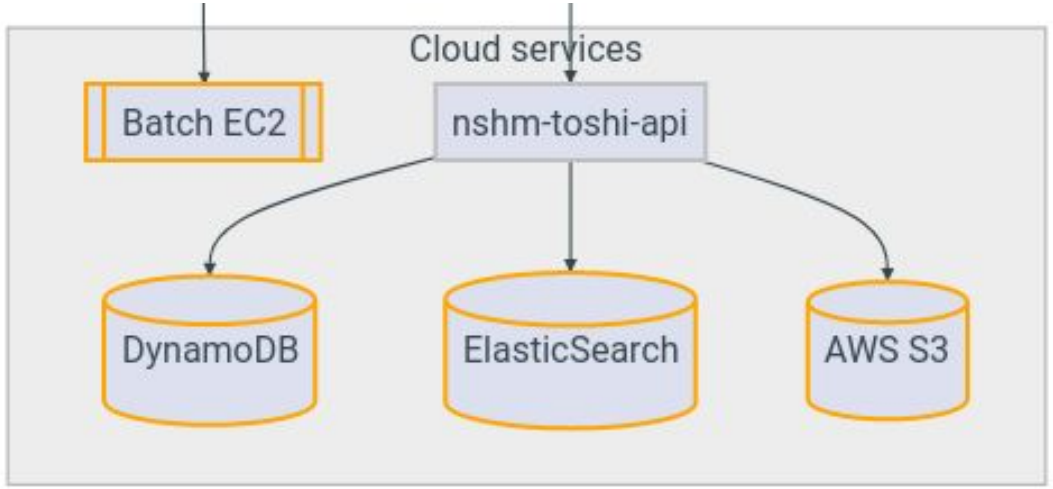


Early days - top 3 Architecture Aspirations/Goals



#1: Do the work, run it where it's most cost/time efficient.

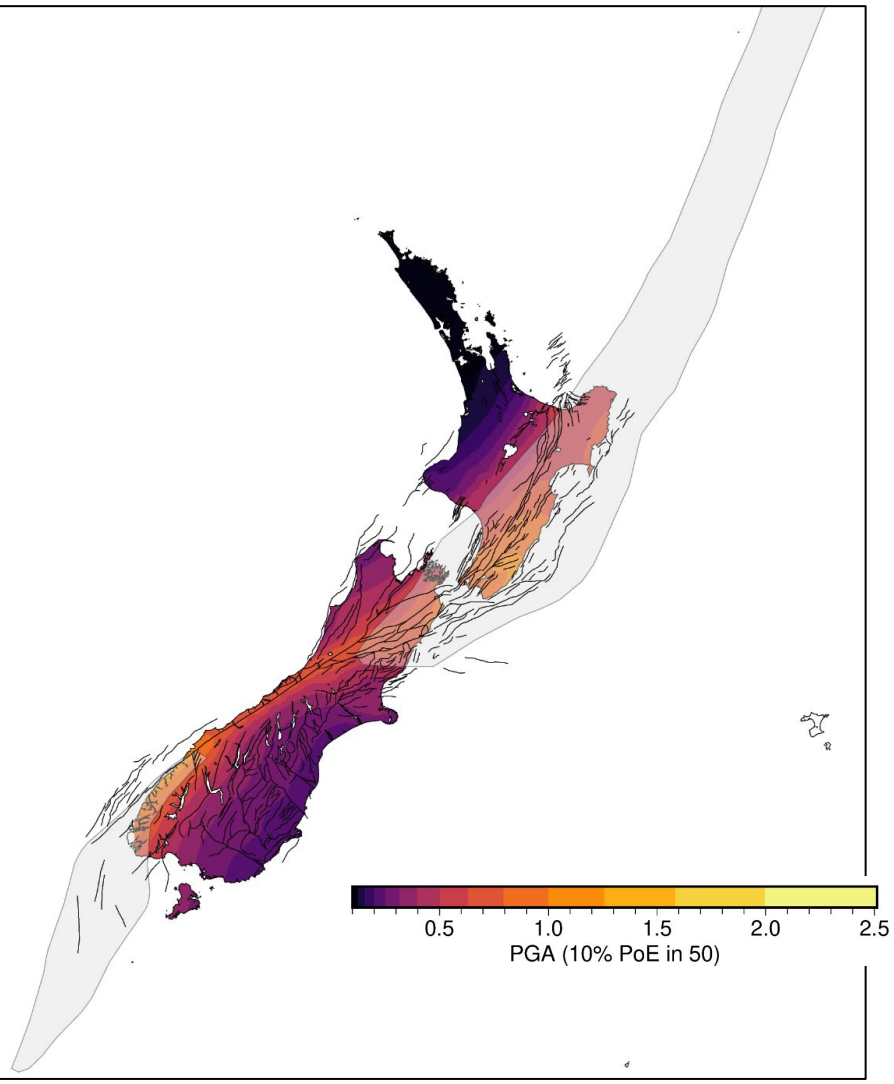
#2: capture all the things(TM) make them accessible



#3: manage costs - both fixed and dynamic.

Hazard modelling at scale

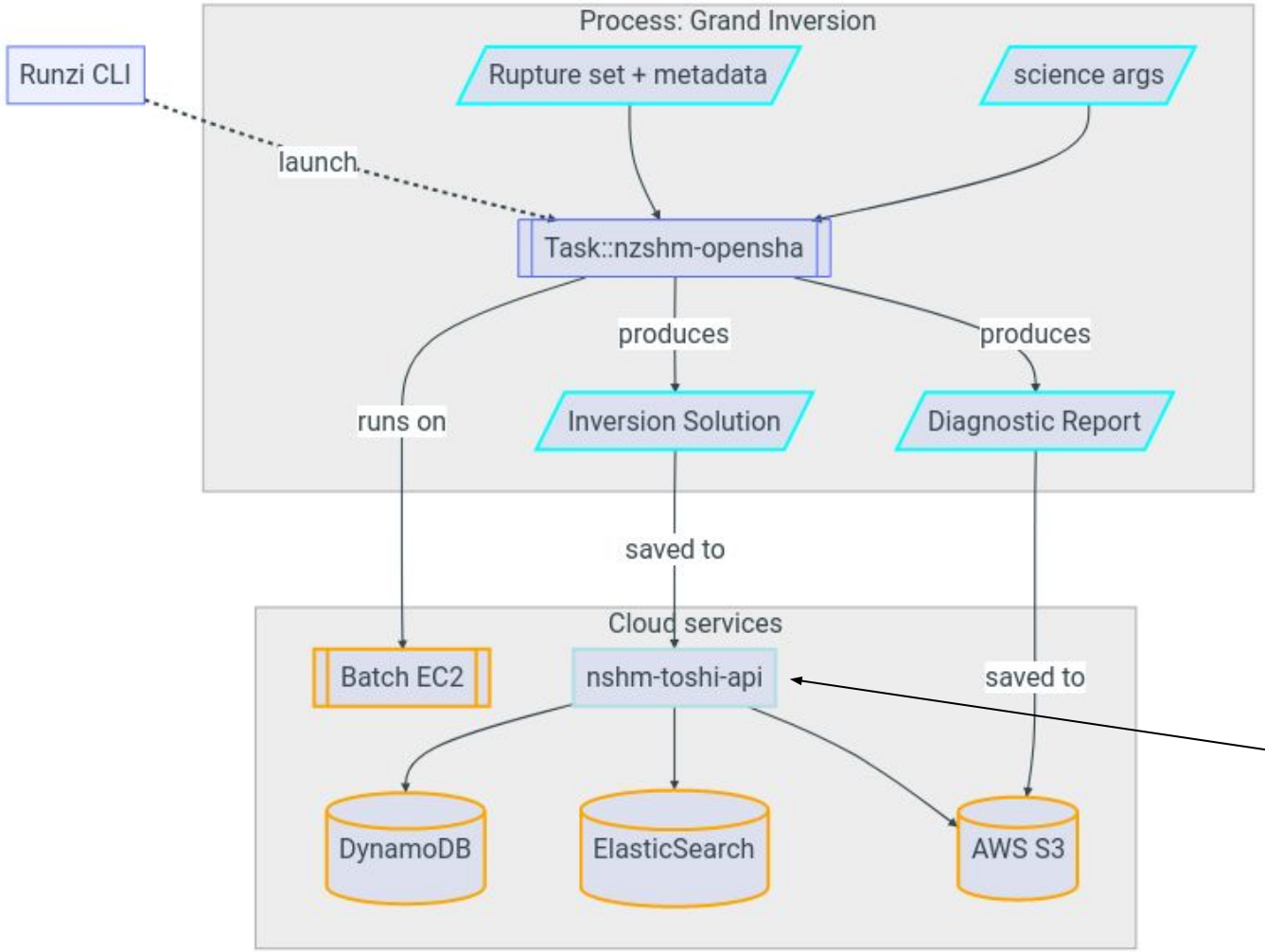
The NZ NSHM 2022 needs a large logic tree to better capture epistemic uncertainty ....



- 324 (SRM) x 3024 (GMCM) = 979,776 total branches / hazard realisations
- >1.1 Million sources per source branch
- 3774 sites (0.1 deg calculation grid)
- 16 spectral periods
- 18 site conditions (Vs30)

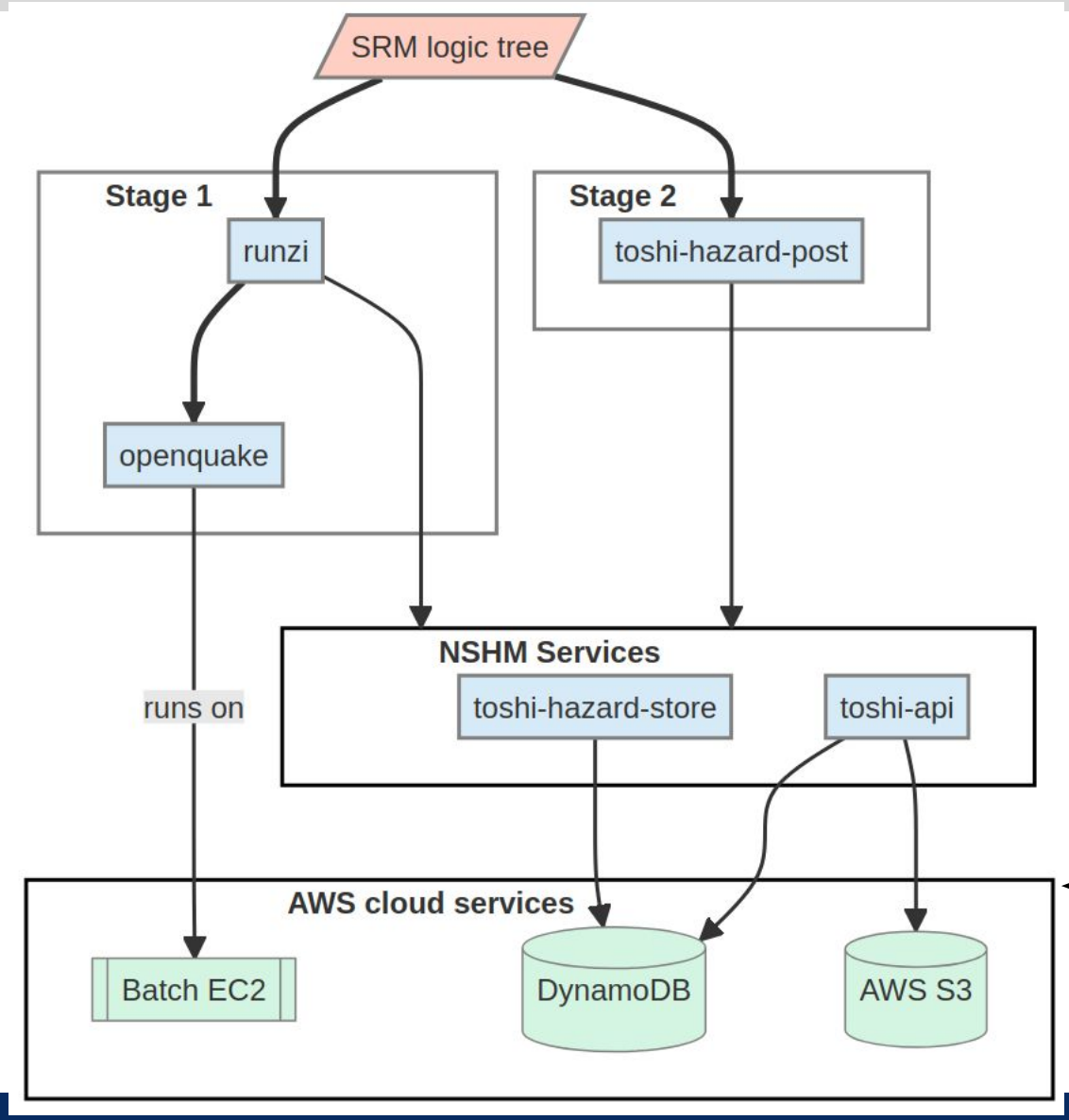
This cannot be run as a single openquake job due to memory and time constraints.

Seismic Event Rate models using Grand Inversion



TUI Web Application

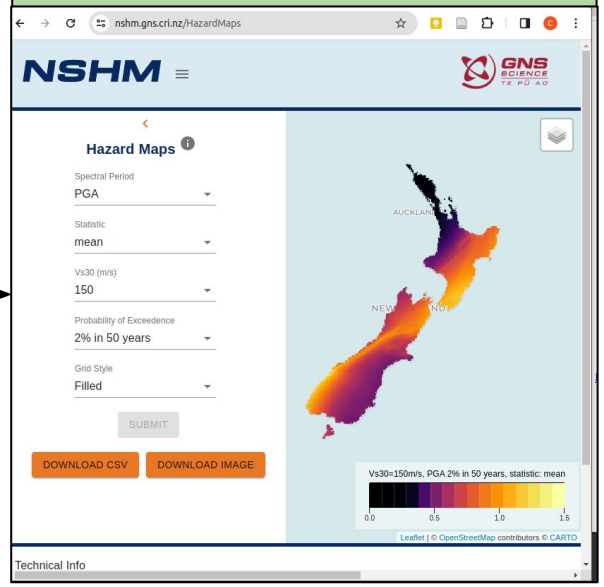
Calculating the Hazard Curves: Working with Large NSHM



The public app sees the same data as the research group via the KororaaAPI

kororaa-api

Kororā Web Application



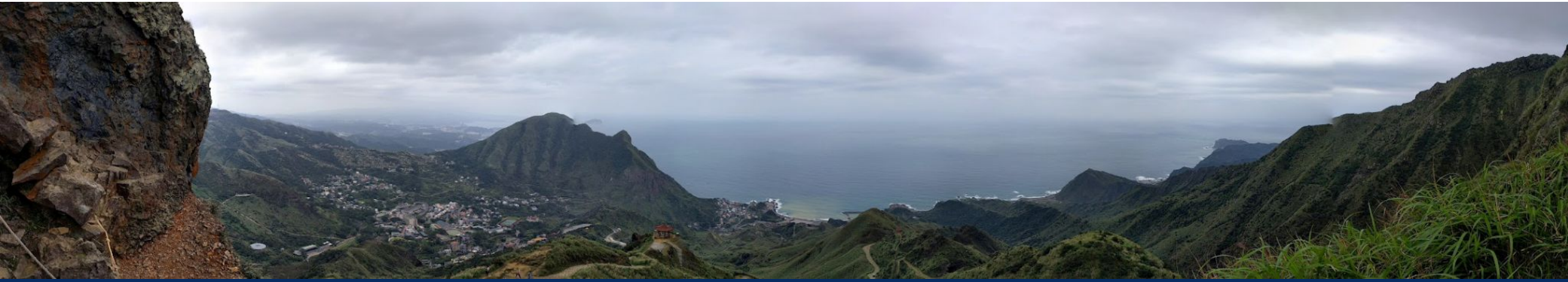
NSHM compute metrics

- 12,500 **Grand Inversion** tasks (~20 compute/months); from 30-120m each
- 2 million **Seismic Hazard** tasks (~175 compute/months); from 5m to 1day each
- 1 scientific paper on hazard calculation with large logic trees

Christopher J. DiCaprio, Chris B. Chamberlain, Sanjay S. Bora, Brendon A. Bradley, Matthew C. Gerstenberger, Anne M. Hulsey, Pablo Iturrieta, Marco Pagani, Michele Simionato; Calculation of National Seismic Hazard Models with Large Logic Trees: Application to the NZ NSHM 2022. *Seismological Research Letters* 2023; doi: https://doi.org/10.1785/0220230226

Next Steps

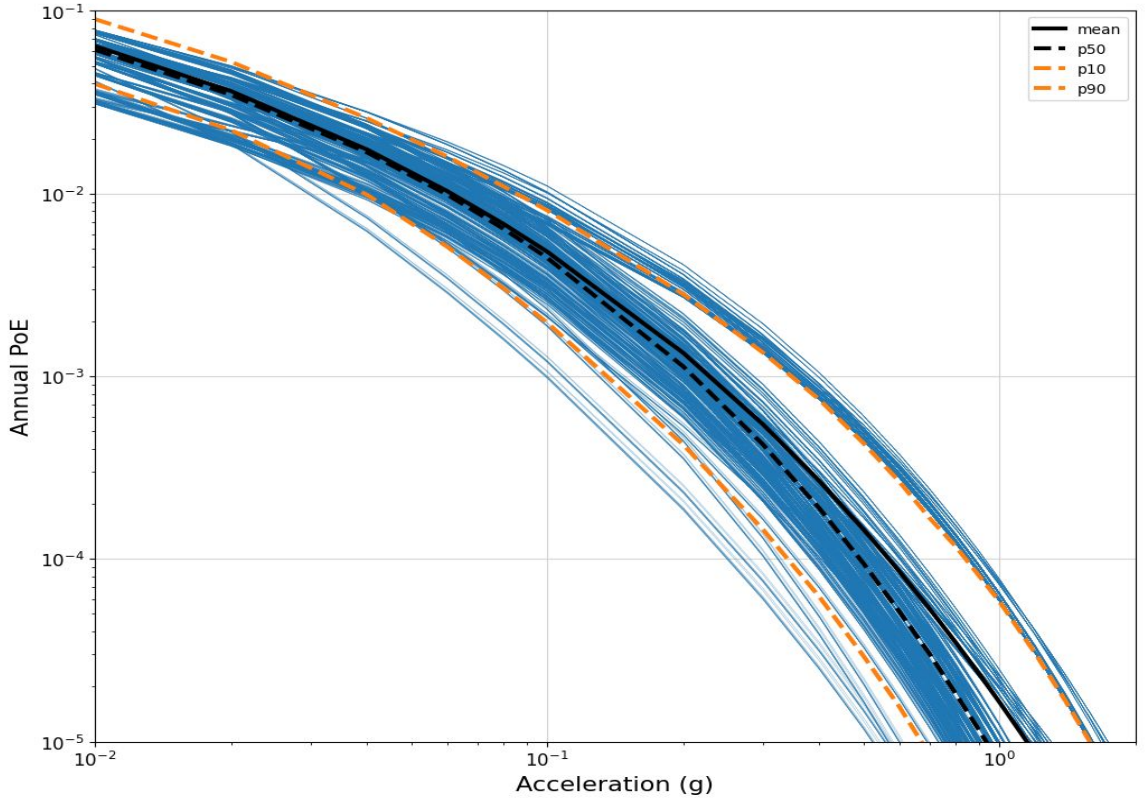
- **Work with GEM** to ensure this method remains compatible with openquake.
- **Hazard on demand**
- **Make code available to public**
  - refactoring and documentation
  - workflow options for cloud and standalone databases



The Logic-tree decomposition...

We now subdivide hazard computations by SRM/GMCM branch permutations. We use cloud compute or HPC, running standard GEM openquake.

Outputs are stored as **HazardRealisation** objects in a Cloud database (AWS DynamoDB).



Comprehending the challenge

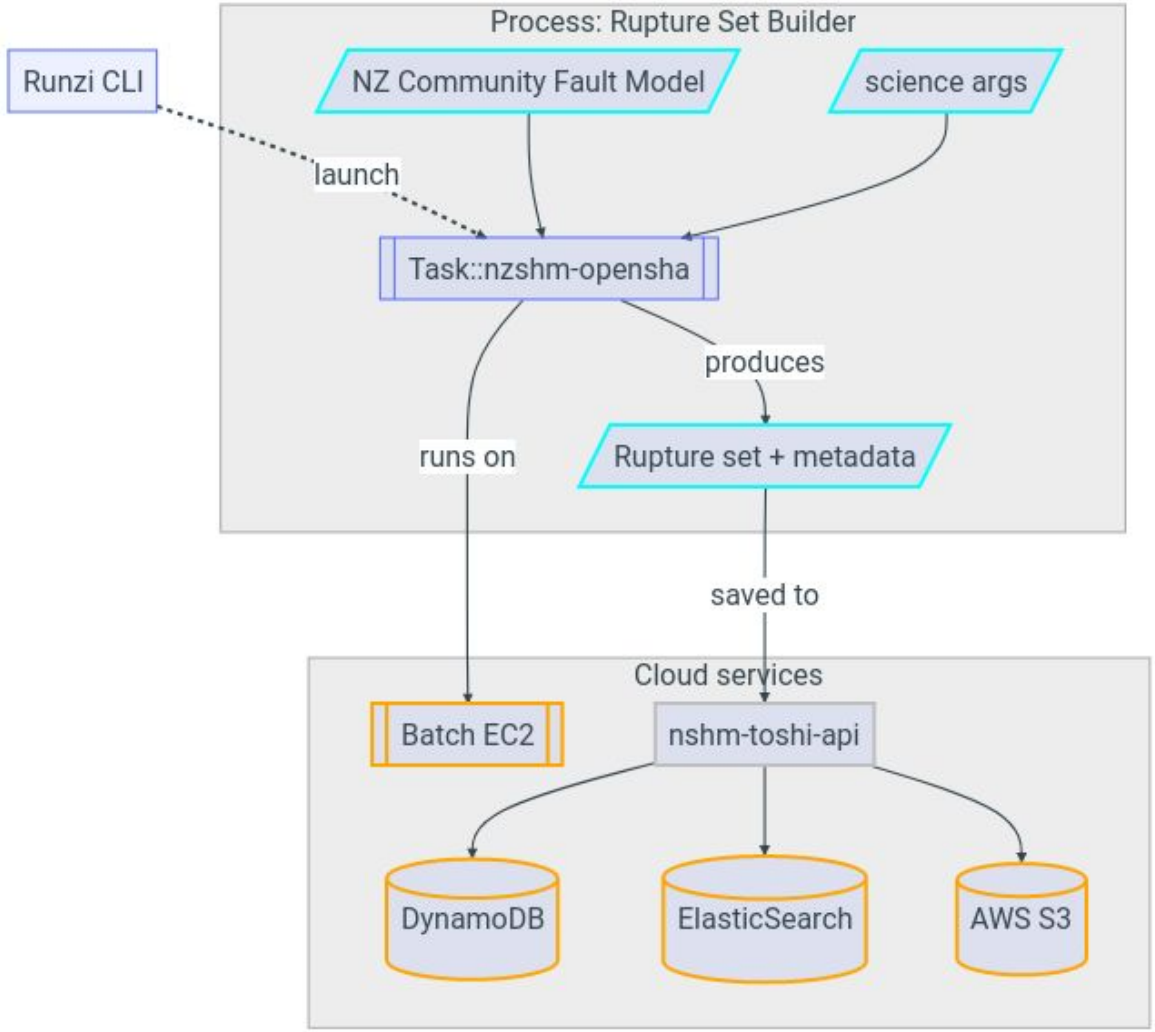
Guiding Principles

- **identify and curate the 'component parts' of the NSHM**
- **make NSHM results traceable and reproducible**

Initially, we didn't understand...

- what is a National Seismic Hazard Model
- The domain language, acronyms, weird jokes
- how the teams and sub-models would fit together
- what science code would need further development
- what IT skills would be needed i.e. dev languages
- scale of the compute resources

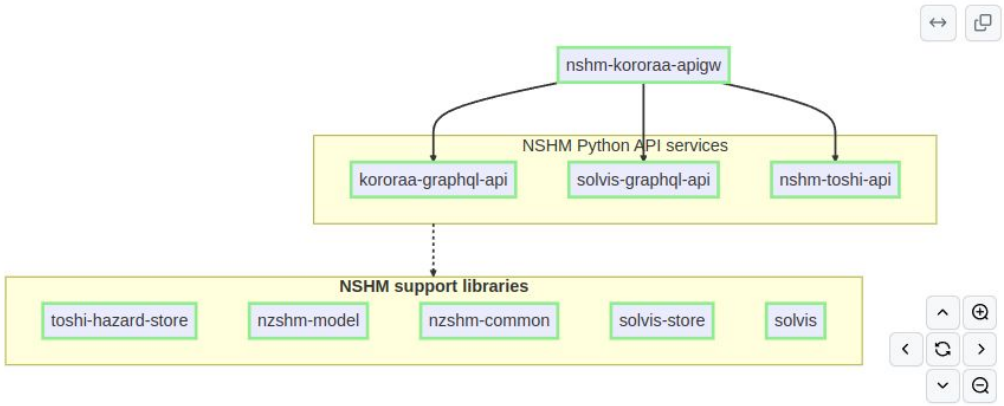
Seismic Event Rate models: Building Rupture sets



CBC WIP. architectural / Process view....

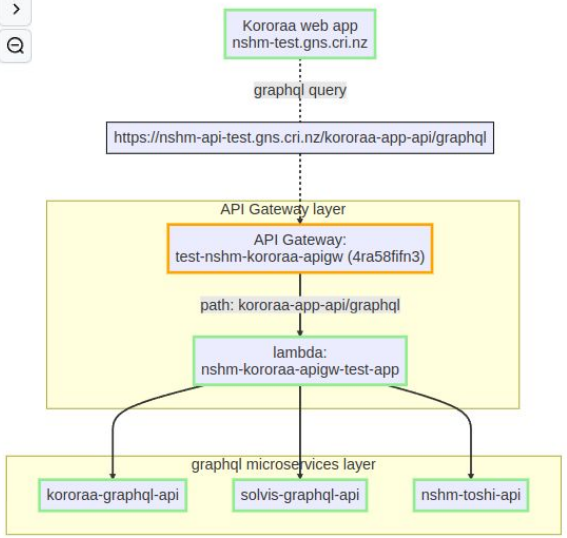
NSHM web service API overview

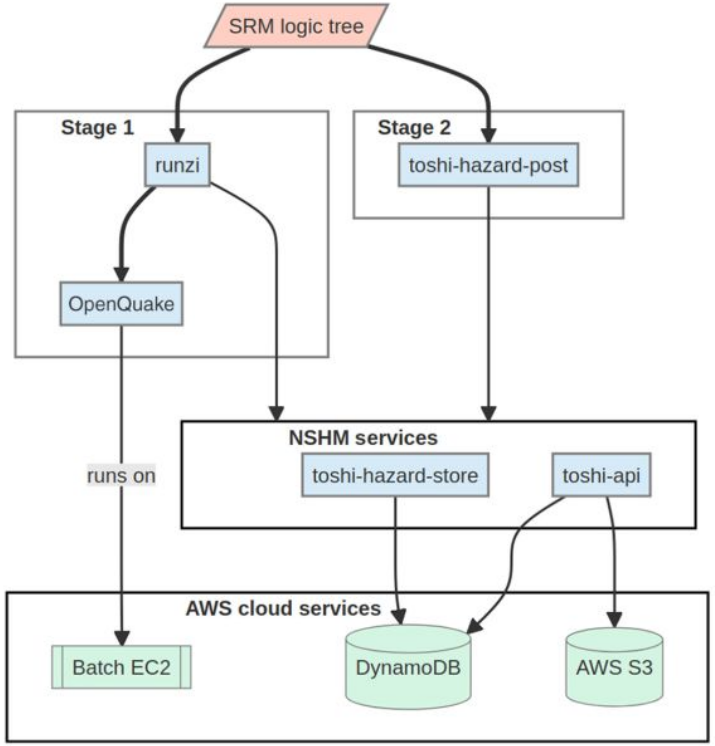
Web service API for the NSHM are graphql APIs, providing flexibility and control for the consumer over what information the API returns. Graphql APIs include a type system, built in documentation and standard error handling.



on API gateway pattern

or the main NSHM web application (Kororaa) . The pattern is commonly known as a
consists of domain-specific services that are then composed into collections by an AI
ommunicate with a single API gateway, which in turn passes on call to the relevant AI
t this style of services organisation.





The Hazard calculation backstory - April 2022

With only a few weeks remaining to keep project on schedule, it became clear that we had a serious problem computing hazard curves in openquake - the chosen PSHA software from GEM.

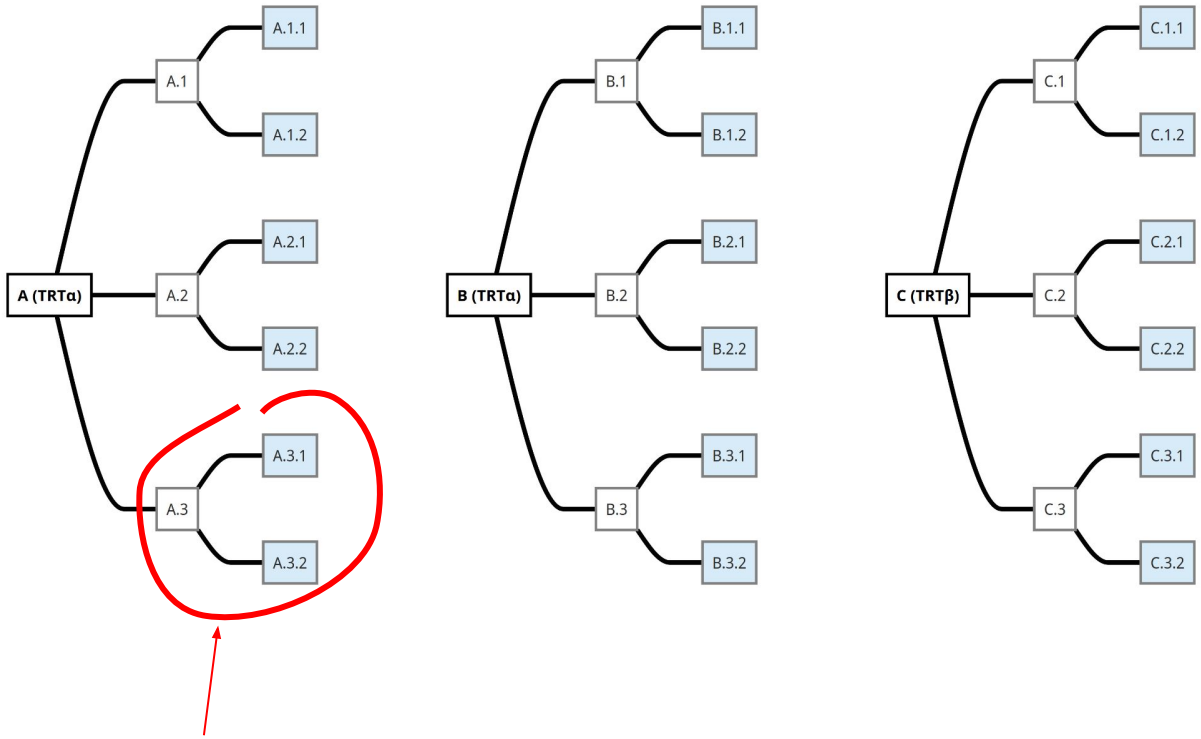
The project philosophy ‘convey the uncertainty’ dictated that we reflect the model diversity through to the hazard calculations. This took the NZ calculation well beyond the design intent of openquake. *Hitting time & memory bounds, talking to GEM, are we doing it right, alternate options like change the philosophy??*

So we explored the concept of splitting the hazard curve calculation and aggregation steps.

If we could do curve aggregation outside of openquake ‘in post’, then we could break down our hazard calcs into many smaller chunks and run them concurrently. These smaller calcs are able to hit the sweet-spot of openquake and complete in realistic time.

It turns out that is actually feasible, and it allowed us to parallel process hazard curves on modest compute nodes ... sometimes up to 750 concurrent jobs, each between 1 to 24 hours duration.

Calculating the Hazard Curves: Working with Large NSHM



- It is possible to calculate independent sub-branches of the logic trees
- Parallelizes and scales across multiple dimensions of the model

The NSHM is not a single forecast of the future but the aggregation of 979,776 weighted models (i.e. possible futures)

These models are represented by a collection of "logic trees."

Forming all combinations of logic tree paths gives the full range of possible models.

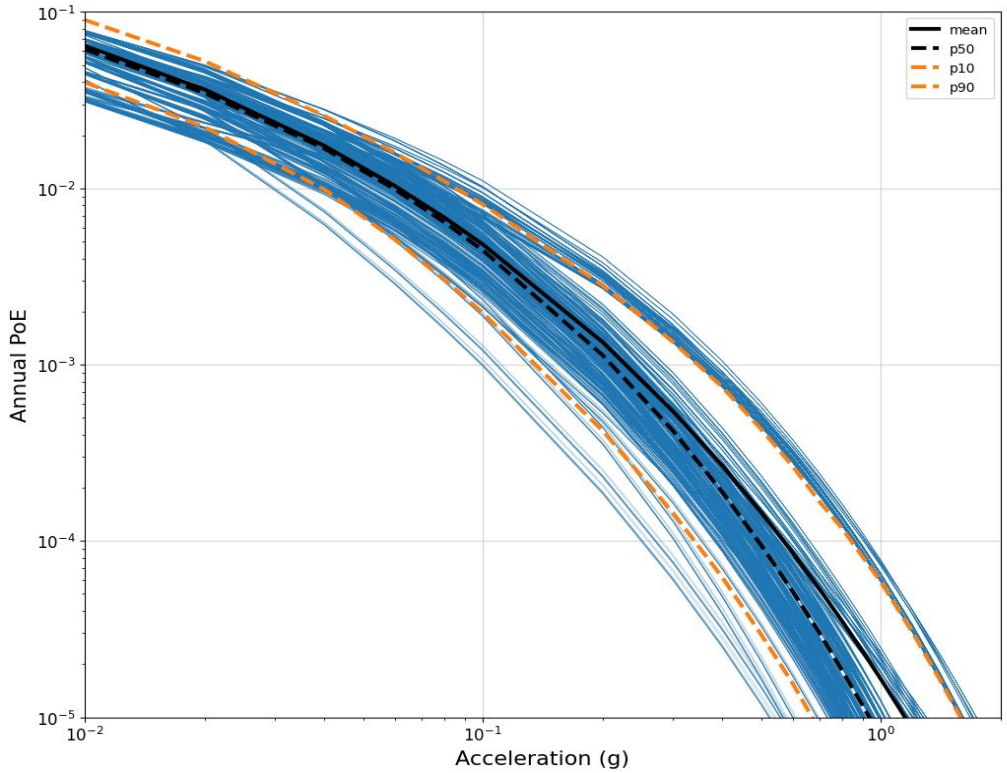
Each model must be calculated for a range of parameters and locations

It is not possible to use conventional seismic hazard software on such a model

DiCaprio, et. al., Calculation of National Seismic Hazard Models with Large Logic Trees: Application to the NZ NSHM 2022. Seismological Research Letters 2023 https://doi.org/10.1785/0220230226

Deconstructing the Problem and Distributing Computation

- The 979,776 models can be decomposed into 912 separate, independent parts
- The 912 computations are distributed onto a cloud compute environment (**AWS Batch EC2**)
- Outputs are stored as HazardRealisation objects in a Cloud database (**AWS DynamoDB**).
- The stored hazard realisations are combined to form all logical paths from which weighted aggregate statistics can be calculated
- These aggregated curves and their dimensions (site, Vs30, etc) make up the ‘final product’



The Hazard calculation: fear and scaling

Now with the stored hazard realisations, weighted fractiles, mean and COV are easily calculated from the set of realisations**.** These aggregated curves and their dimensions (site, Vs30, etc) make up the ‘final product’ i.e **NZ NSHM_v1.0.4**

Advantages: stored branch realisations provide benefits for experimentation and further improvements:

- Parallelizes and scales across multiple dimensions of the model
- New logic tree configurations
- easy sensitivity testing
- sub-sampling to find minimum-viable logic trees
- minimise calculation overheads
- easy branch reweighting

DiCaprio, et. al., Calculation of National Seismic Hazard Models with Large Logic Trees: Application to the NZ NSHM 2022. Seismological Research Letters 2023 https://doi.org/10.1785/0220230226

Our team: Computational Working Group

How:

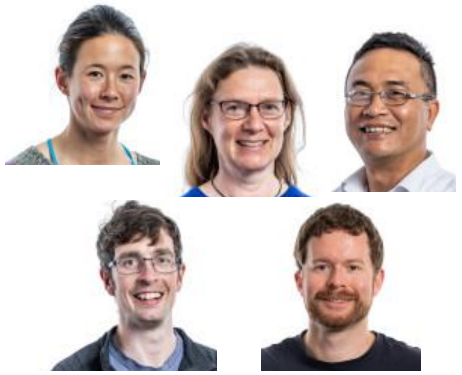
- Start with 1 PM + 1 Architect + 1 Scientist + 1 Subject Expert
- Develop relationships with outside collaborators (GEM, USGS)
- Support and encourage project scientists to contribute code
- ‘Borrow’ experienced Java dev from GNS apps team
- Hire junior devs for Web UI’s, starting them early to build capability
- Broad mix of skills and experience (IT/science/industry)
- Everyone on team doing hands-on dev (science-devops)

What/where:

- Many locations + all the covid lockdowns
- Slacked the communication
- Gitted all the code
- Zoom-Slack-Teamed the meetings

Computational Working Group

SRM Working Group



GMCM Working Group

Public web site: nshm.gns.cri.nz (aka Kororā)