

VEDANT PURI

<https://vpuri3.github.io/>

SUMMARY

Builds scalable transformer architectures grounded in numerical analysis and scientific computing. Developed FLARE (Fast Low-Rank Attention Routing Engine), scaling attention to million-token regimes on a single GPU via Triton. Background includes numerical PDE methods, distributed training, and PyTorch systems implementation.

EDUCATION

Carnegie Mellon University (CMU) <i>Ph.D Mechanical Engineering.</i> Advisors: Levent Burak Kara, Yongjie Jessica Zhang	Jan 2022–Present
University of Illinois Urbana-Champaign (UIUC) <i>B.S. Engineering Mechanics, B.S. Mathematics.</i>	2015–2019

SELECTED RESEARCH CONTRIBUTIONS

FLARE: Fast Low-Rank Attention Routing Engine <i>Efficient attention architectures</i>	2025 (Under review)
• Reduced memory from $\mathcal{O}(N^2)$ to $\mathcal{O}(NM)$. • Implemented Triton + PyTorch modules with reproducible scaling results. • Evaluated on PDE surrogate, NLP, and vision tasks. • arXiv: 2508.12594, code: vpuri3/FLARE.py.	

Equation-based PDE modeling with neural fields <i>Hybrid data + physics methods</i>	JCP 2025
• Neural fields used as nonlinear spatial ansatz in reduced-order modeling. • Physics-based evolution combined with learned spatial representations.	

EXPERIENCE

Carnegie Mellon University <i>Research Assistant</i>	Jan 2022–Present
• Phase field simulations of lithium dendritic growth in solid-state batteries. • Turbulence closure modeling with differentiable physics.	
Julia Computing <i>Intern Engineer</i>	Apr 2021–Nov 2021
• Built numerical solvers for scientific machine learning ecosystem in Julia.	
Carnegie Mellon University <i>Research Assistant</i>	Sep 2020–Jan 2021
• Developed differentiable geometry representations and meshing algorithms.	
Argonne National Laboratory <i>Research Assistant</i>	Mar 2020–Sep 2020
• Executed large-scale simulations of turbulent airflow over urban landscapes on supercomputers.	
Argonne National Laboratory <i>Research Assistant</i>	May 2018–Jul 2018
• Executed high-fidelity fluid dynamics simulations and analyzed turbulent statistics for closure modeling.	
National Center for Supercomputing Applications <i>Intern</i>	Sep 2017–May 2018
• Numerical simulation of spacetime metric for gravitational wave simulations.	

TEACHING

Carnegie Mellon University <i>Teaching Assistant, Numerical Analysis</i>	Spring 2025
Carnegie Mellon University <i>Teaching Assistant, Discrete Differential Geometry</i>	Spring 2023
University of Illinois <i>Course Assistant, Introductory Statics</i>	Spring 2016–Fall 2017

PUBLICATIONS

- Puri, V. et al., *FLARE: Fast Low-Rank Attention Routing Engine*. arXiv 2025. (Under review)
Puri, V. et al., *SNF-ROM: Projection-based nonlinear reduced order modeling with smooth neural fields*. JCP 2025.
Shankar, V., Puri, V., et al., *Differentiable physics closure modeling for Burgers' turbulence*. MLST 2023.

ACTIVITIES & AWARDS

World Conference on Computational Mechanics <i>Best poster in fluid dynamics</i>	2024
University of Illinois <i>Theoretical and Applied Mechanics Merit Award</i>	2019
Society for Engineering Mechanics, UIUC <i>President</i>	2019

TECHNICAL SKILLS

- Machine Learning Systems:** PyTorch, Triton
Numerical Computing: Linear systems, finite elements, spectral methods
Programming Languages: Python, Julia, C, Fortran 77, MATLAB, UNIX, L^AT_EX